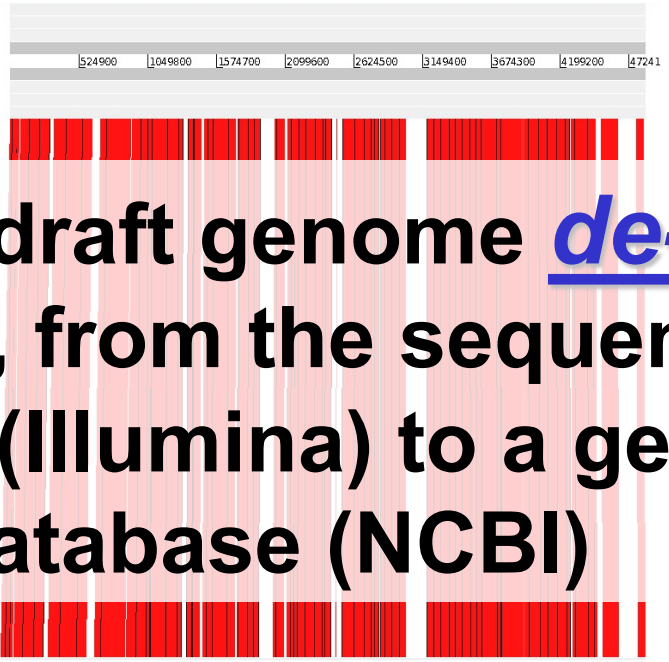
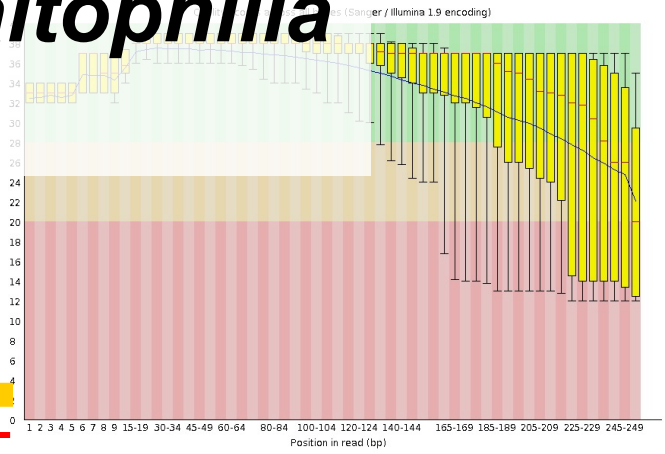
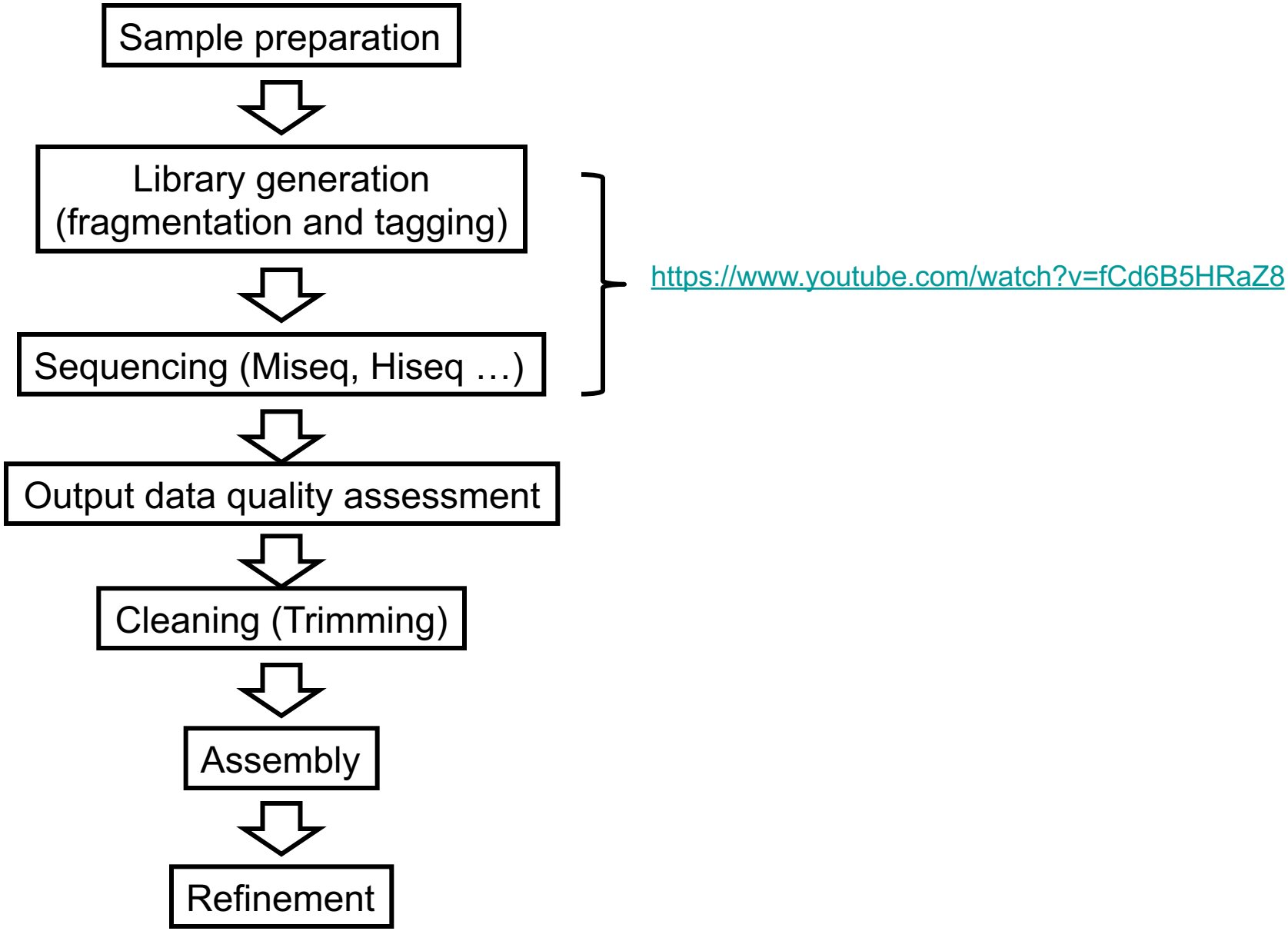


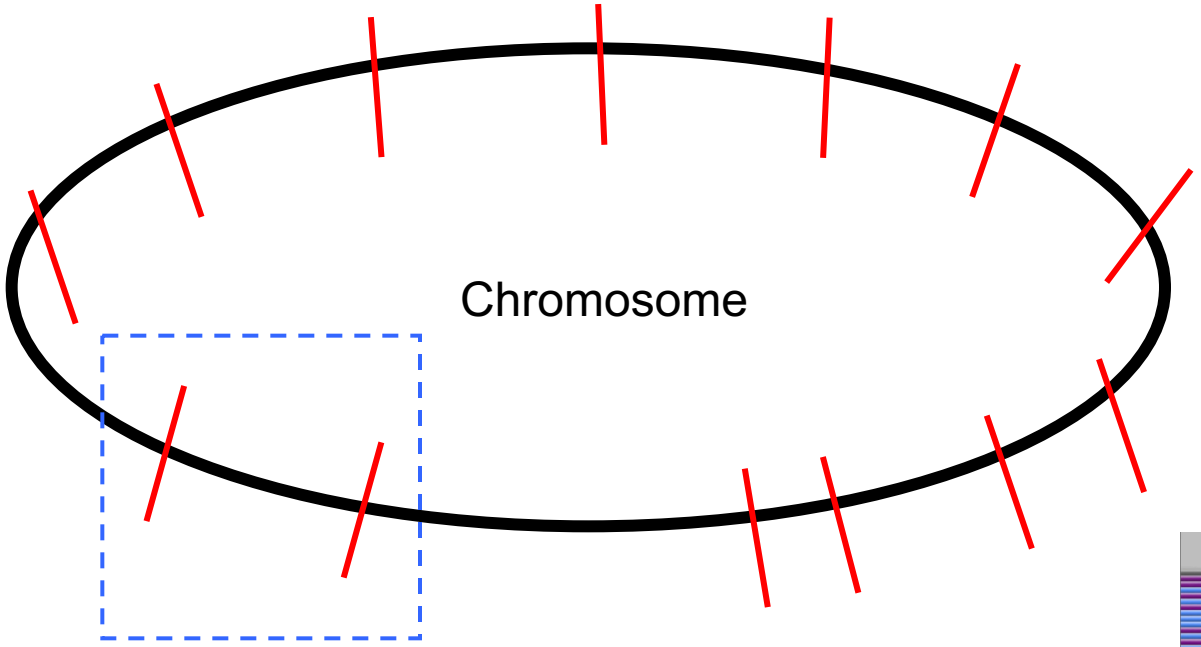
Bacterial draft genome *de-novo* asembly, from the sequencing machine (Illumina) to a genome database (NCBI)

An example case:
Assembly of
Stenotrophomonas maltophilia
strain **PG157**

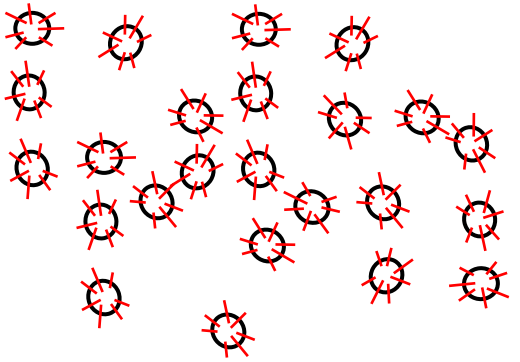


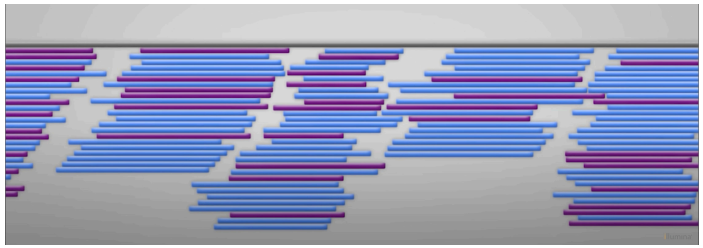
Assembly process with illumina® sequencers

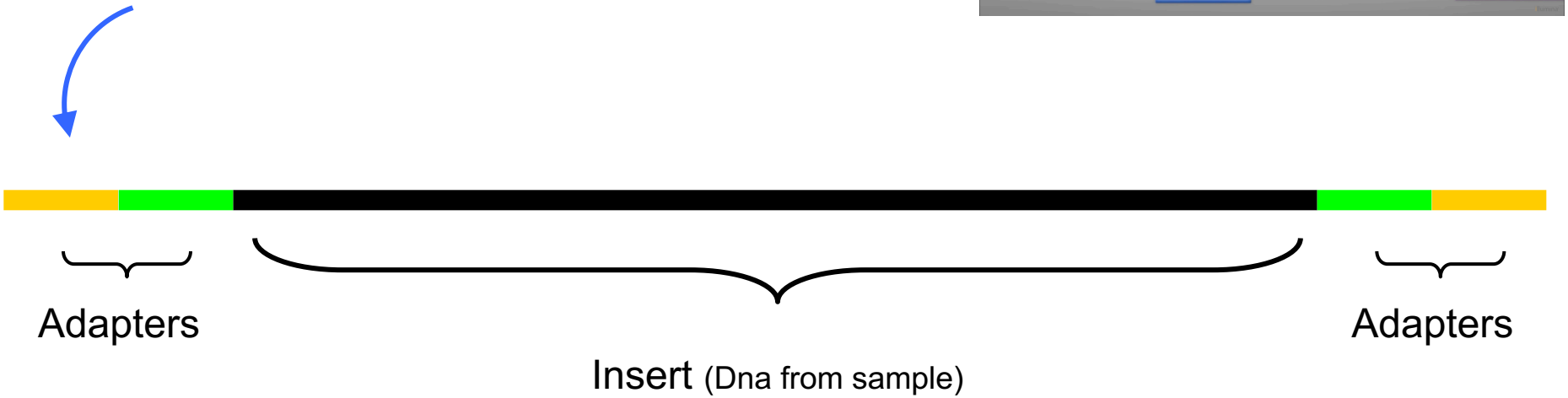




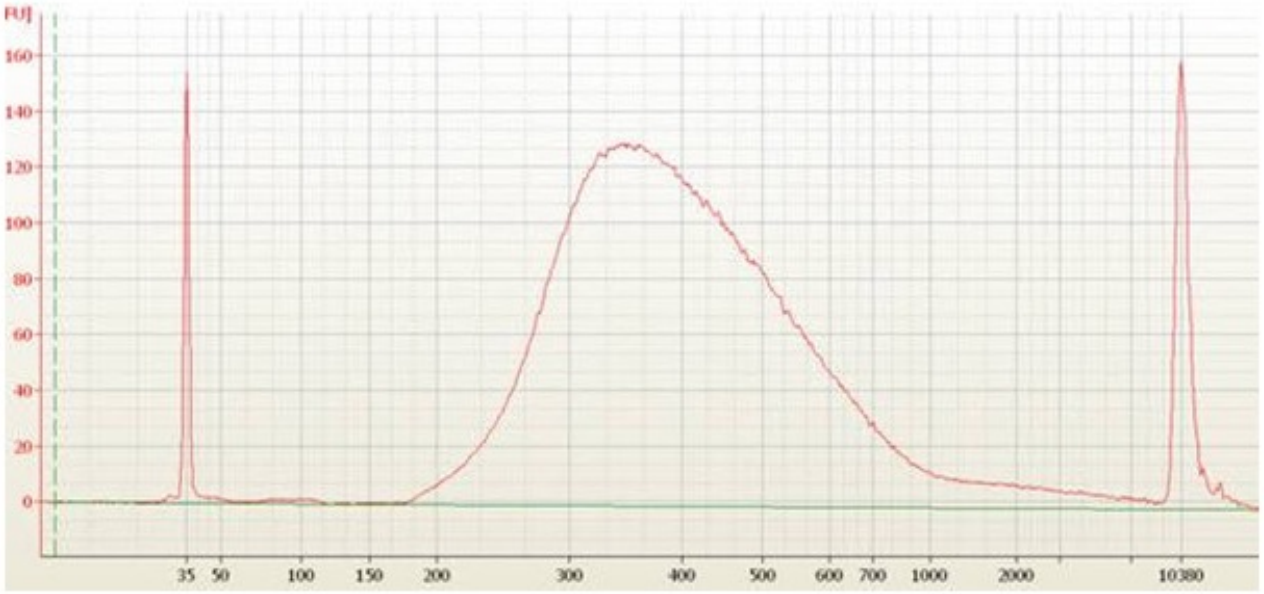
You are “cutting” millions of chromosomes at the same time, each one for different places

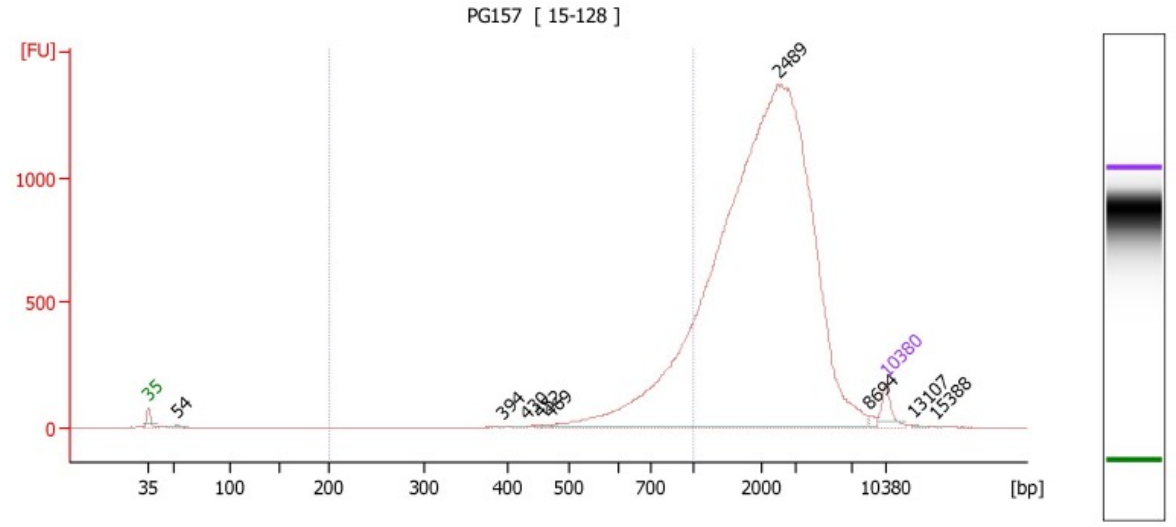






Library Status





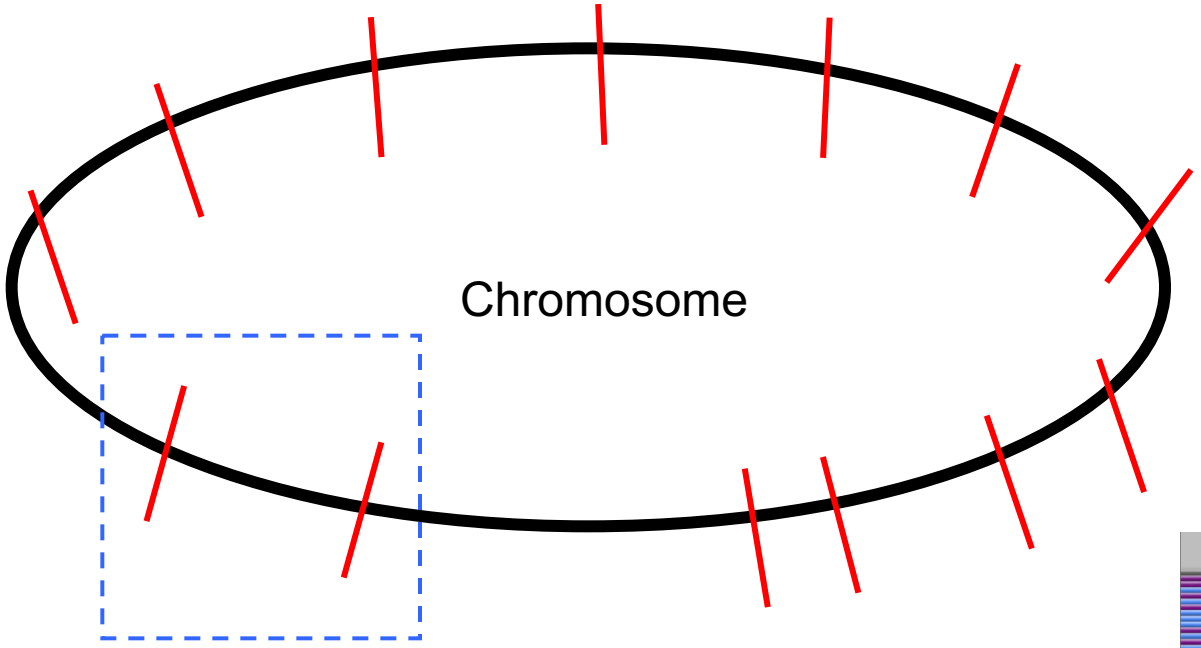
Agilent bioanalyzer

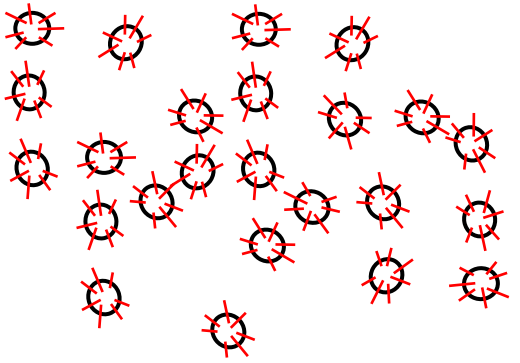
MiSeq

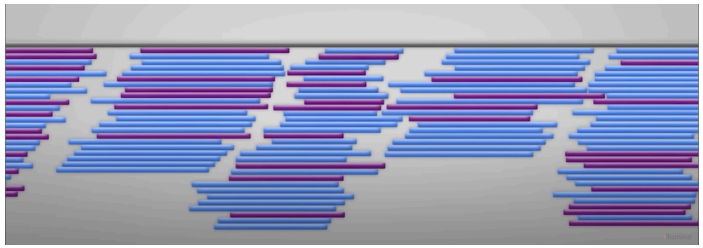
Lanes:	1
Reads / Lane:	25E6
Read size:	300 * 2
Cost per run:	$ ~1.4k
Cost / Gb:	$ 100
Cost (machine)	$ ~ 100k

HiSeq

Lanes:	8
Reads / Lane:	180E6
Read size:	100 * 2
Cost per run:	$ ~25k-30k
Cost / Gb:	$ 30 – 40
Cost (machine)	$ ~ 700k







R1

R2

Short insert: reads include adapters in both ends

Very Short insert: reads include adapters and “invented” data (usually poli A)

Mandatory parameter:

number of cycles => a priory, all reads will have the same size

Sequencer output: fastq file

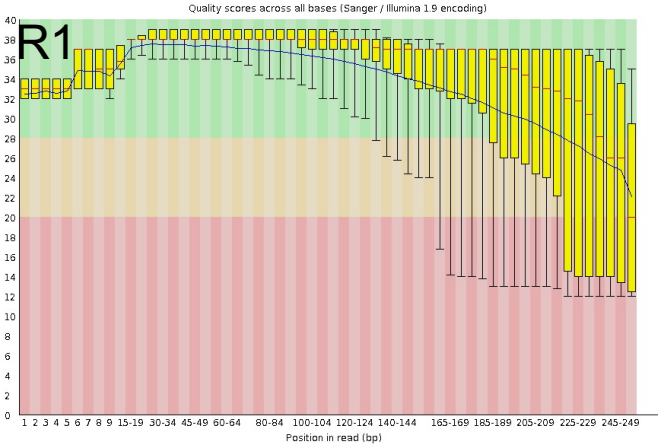
```
@M01269:69:000000000-AD379:1:1101:15470:1360 1:N:0:2
AGGTTGTGTGGCATCACCGTGGCACCAATGATGCCGATCGCGATATACAGCGCGTGCGGATCAGTCACCACCTGCGCGCGC
GGAATGAATCCGCCCAGCACGTCCATCACCGGCGGTGCGGCCAGCGCGATCTGCACCATGAAGCAGCCGAAGATCACCATC
AGCAGCGCGATCACGAAGGCTTCCAGCGCACGGAAGCCGCGGTTCATCAGCAGCAGTACCAGCAGCGTGTCCATGGCGGCG
ATCACCG
+
ABBBBFBBFFBBGGGGGGGGGGGHHGHCGHFHGHHGGGGGGGGGGGHHHHHGGGGGGGGGGGHHHHHHHHHGHHHGGGGGG
GGCGGHHFHHHGGGGGHHEHGHHGGGHHHHG?DFGC:B??@D?EG??BDFFFFFFFF.BFFFFFFFFF;-
@9:BFFFFFFF/B9AE@F=DAFFF.....;9FFFF/B;CBDDFFFDFFFADFD>FFFFFFFFF.;BFFBFBAFFF>DFDB
F//9/;>--:9DEBFBD
@M01269:69:000000000-AD379:1:1101:15606:1380 1:N:0:2
GTACCTGAAGGATGAGTCGAGCCATCCCACCGGCAGCCTGAAGCACCGGCTGGCGCGCTCGCTGTTCCTGTATGCGCTGGC
CAATGGCTGGTTGCGTGAAGGGCGTCCGGTGATCGAGGCCTCCAGCGGCTCGACCGCCGTGTCCGAAGCCTATTTCGCGCG
GCTGCTCGGCCTGCCGTTCATTGCGGTGATCCCGGCCTCCACCTCGCCTGAAAAGATCGCCGCGATCGAGTTCCACGGCGG
CCGTTGC
+
AAAABFFFFFFFGGGFGGGGGGGHHHHHHHGGCGGCGHHHHHHGHHG?AEFGGGGGGGCFGGF?EHHHHHHHHHHGGGG?F
HHHGHHHGGGGHGFCGGGAGFCEGGGGGCFGGHH.A.<-EHHHHEHGC-
@FCCGG?C?D@DFFBDGFF;FFFFFFFF?FA;D;BF?F/D-?DDFFFADFBFFFF/;DA-99BF.A-9BFFEF..;A.-
>D?.99//.;FA..-9-9-.9..9:/99BFA-;-999@A.;
```

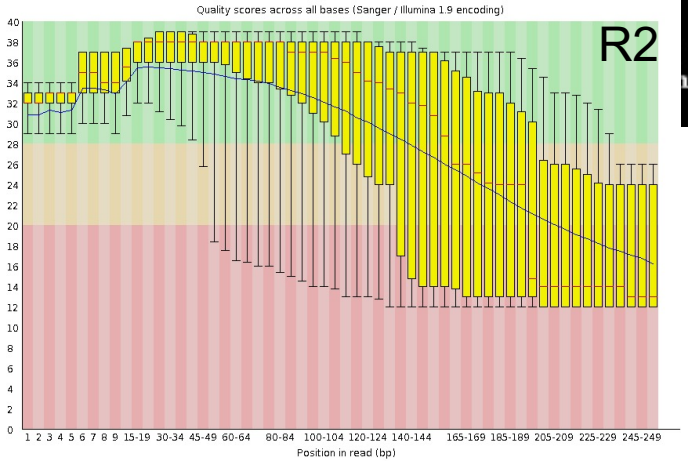
M 2015	PG157_S2_L001_R1_001.fastq.gz	476617
	PG157_S2_L001_R2_001.fastq.gz	476617
D 2014	PG157_S2_L001_R1_001.fastq.gz	71430
	PG157_S2_L001_R2_001.fastq.gz	71430
M 2015	PG157-2_S6_L001_R1_001.fastq.gz	370548
	PG157-2_S6_L001_R2_001.fastq.gz	370548
D 2014	PG157_S7_L001_R1_001.fastq.gz	106550
	PG157_S7_L001_R2_001.fastq.gz	106550
M 2015 EDTA	PG157-EDTA_S9_L001_R1_001.fastq.gz	460114
	PG157-EDTA_S9_L001_R2_001.fastq.gz	460114

FastQC

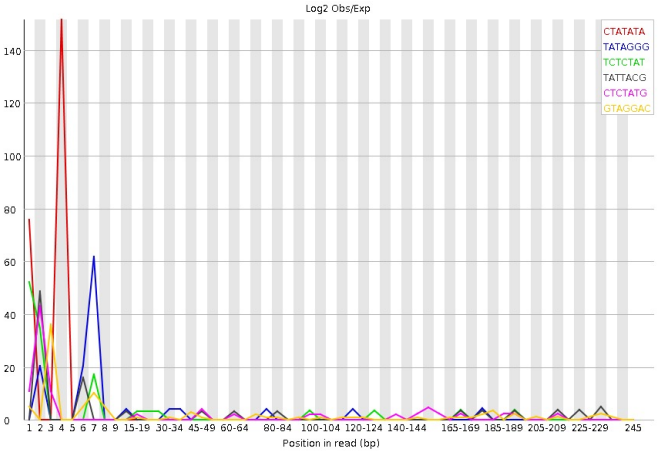
A quality control tool for high throughput sequence data.

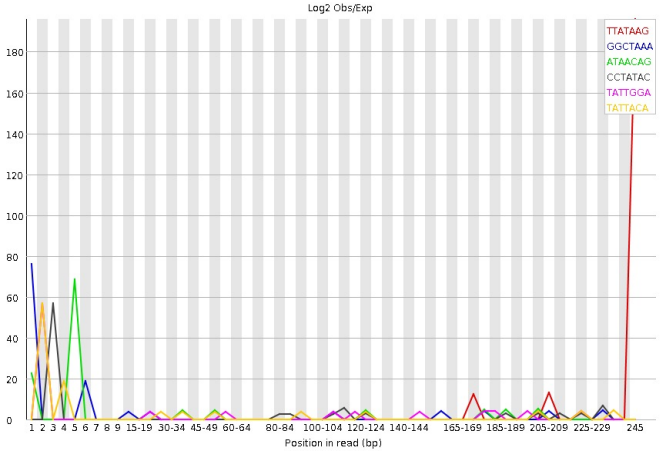
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

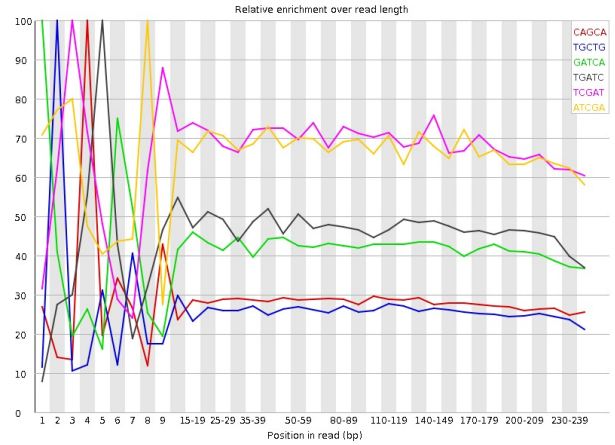


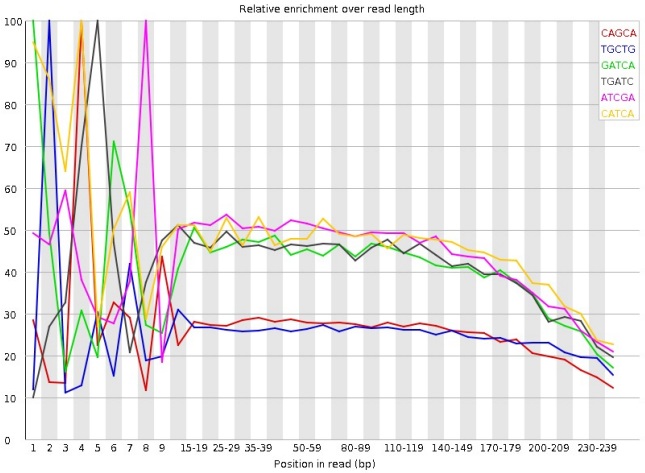


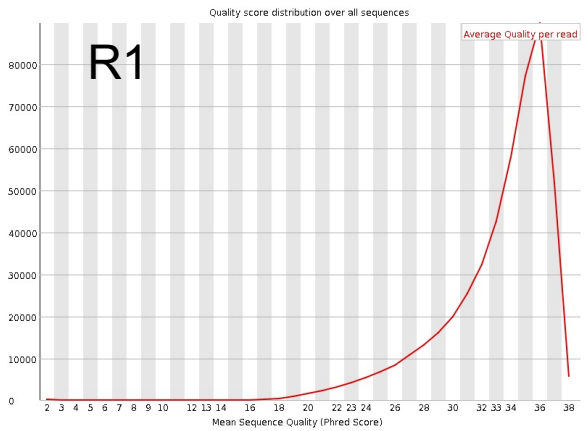
FastQC
(MiSeq PG157)



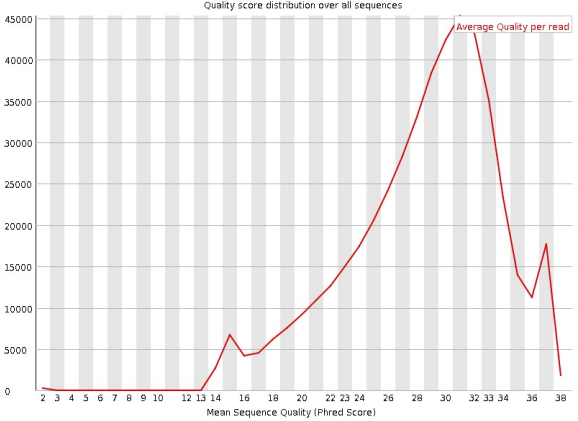


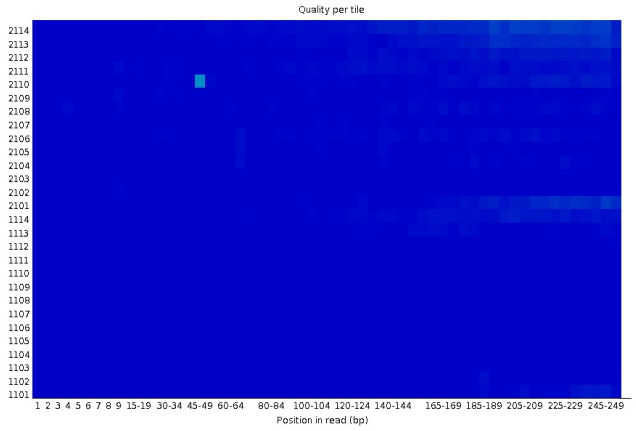


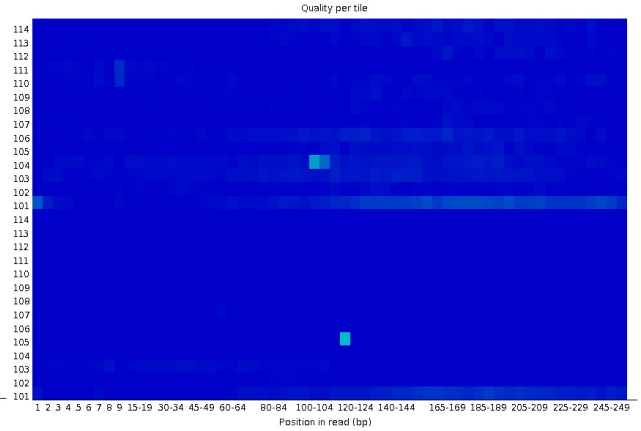


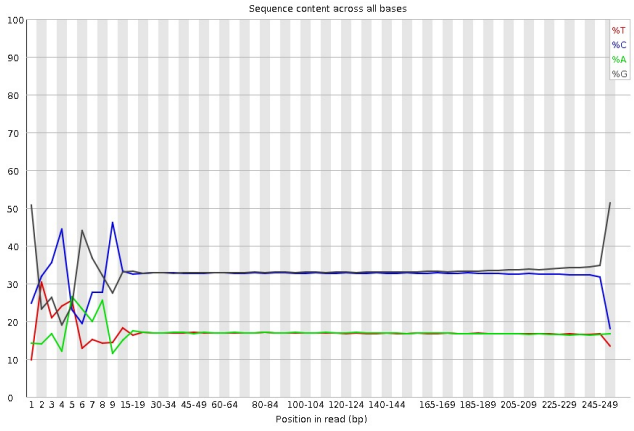


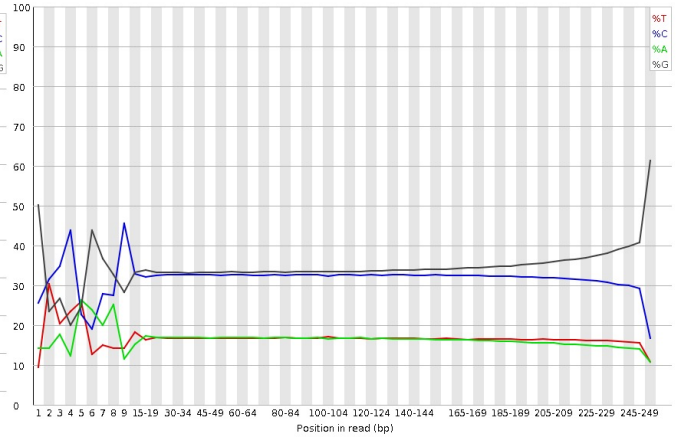
R2



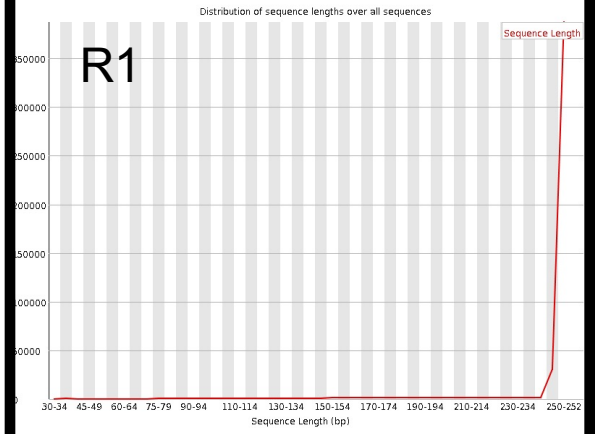


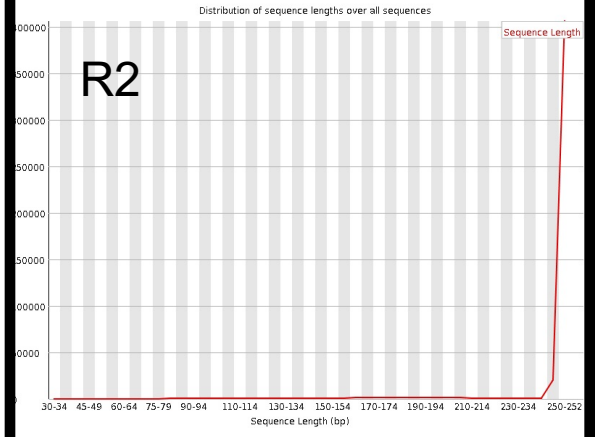


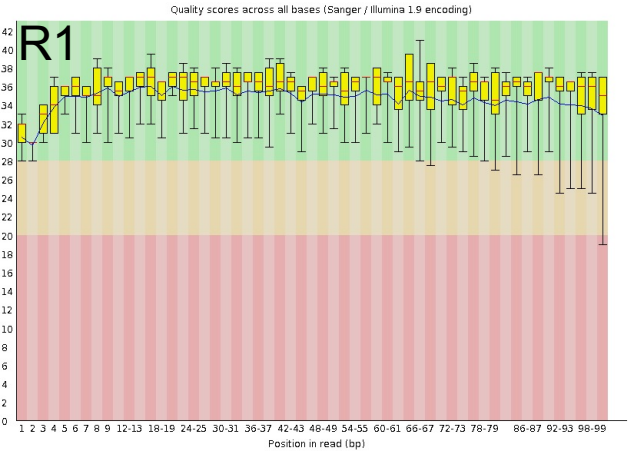


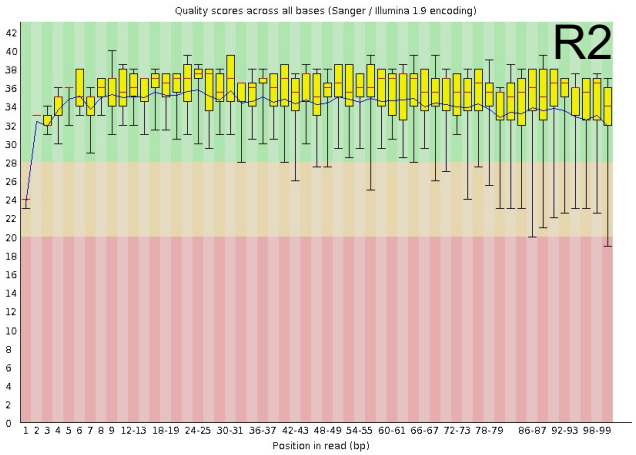


MiSeq: performs a trimming





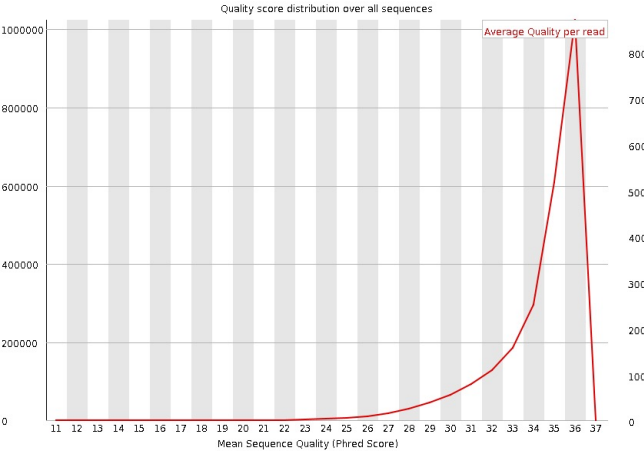


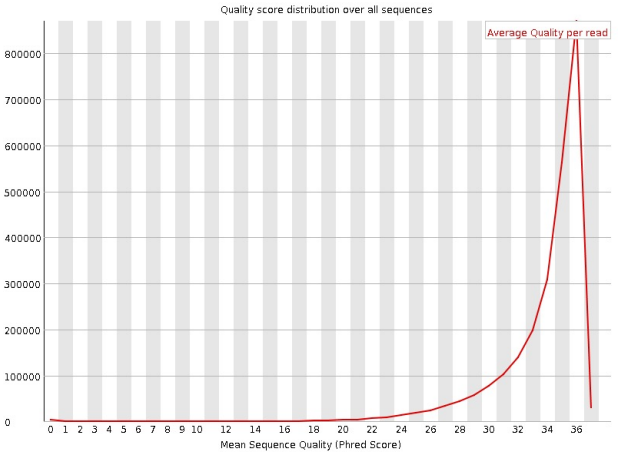


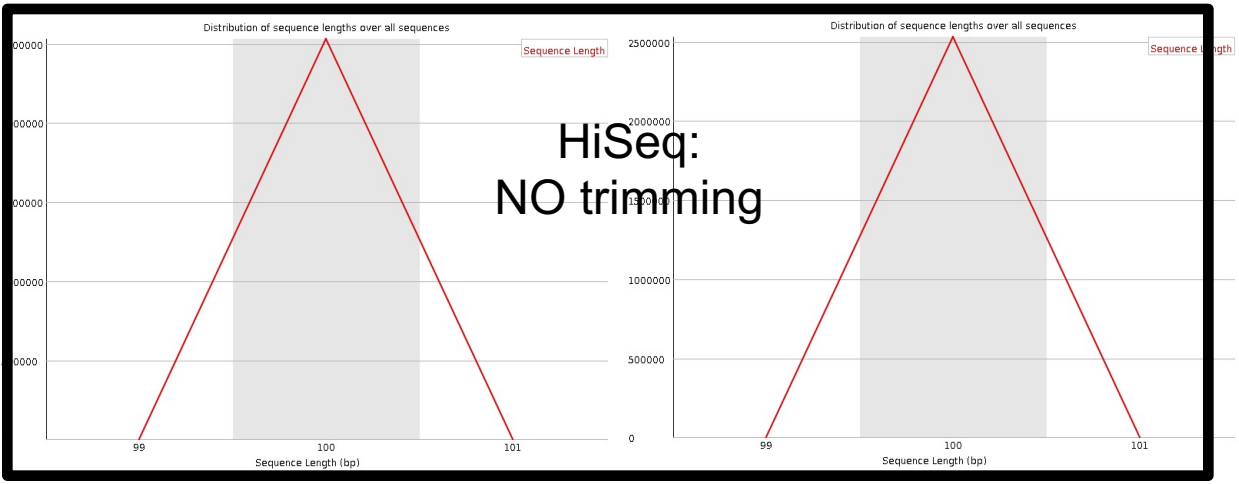
FastQC

(HiSeq ERR350126)

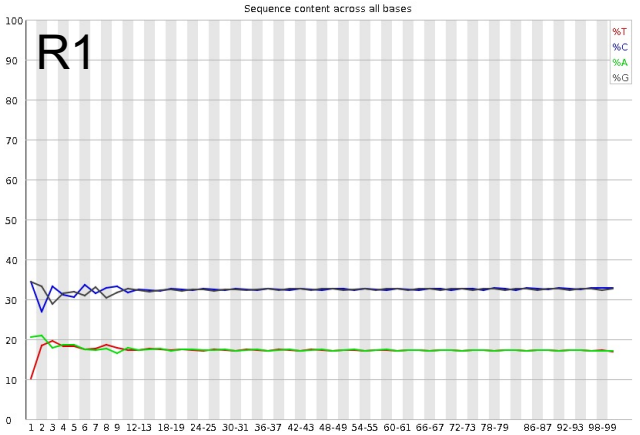
Number of reads: 2529915

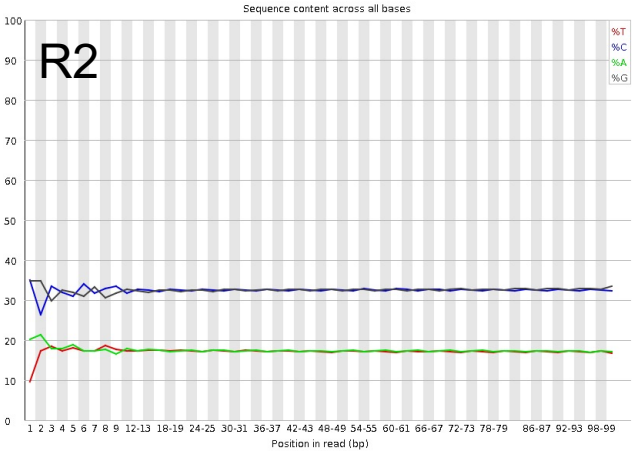


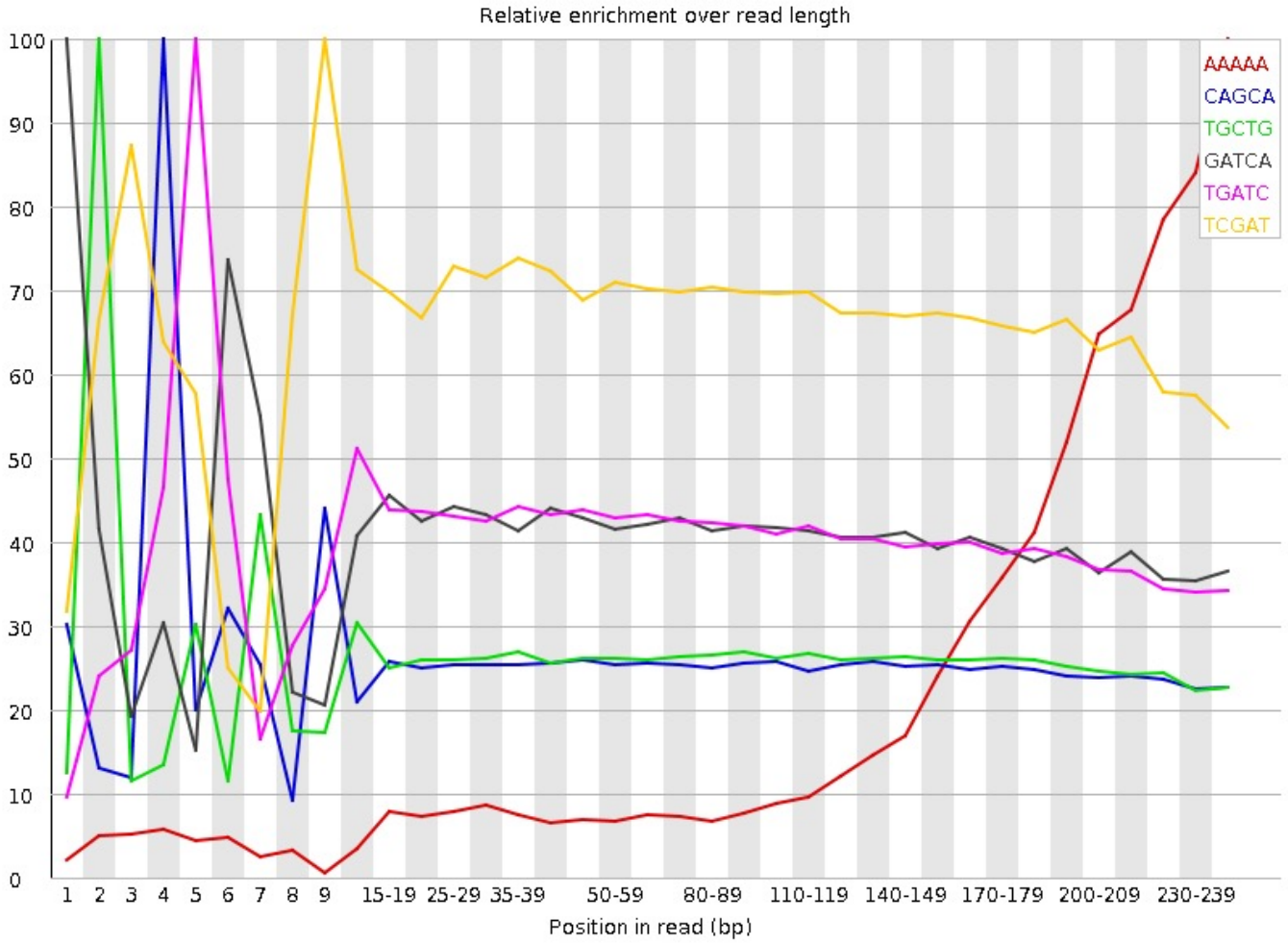




HiSeq:
NO trimming






Relative enrichment over read length
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 15-19 25-29 35-39 50-59 80-89 110-119 140-149 170-179 200-209 230-239
Position in read (bp)
AAAAA
CAGCA
TGCTG
GATCA
TGATC
TCGAT

Good insert: reads include adapters in 5’ end

R1 R2

Short insert: reads include adapters in both ends

Very Short insert: reads include adapters and “invented” data (usually poli A)

Adapter cleaning

- **Univec** database (only illumina adapters)
- UCSC **blat** (vecscreen uses too much RAM)

Fast heuristic sequence search tool (searches for similar sequences)
http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/blat/
http://hgdownload.cse.ucsc.edu/admin/jksrc.zip
https://genome.ucsc.edu/FAQ/FAQblat.html
http://www.ncbi.nlm.nih.gov/pubmed/11932250

- **Skewer**

A fast and accurate adapter trimmer for next-generation sequencing paired-end reads
http://www.ncbi.nlm.nih.gov/pubmed/24925680/
http://sourceforge.net/projects/skewer/

- **Cutadapt**

Removes adapter sequences from high-throughput sequencing reads
http://dx.doi.org/10.14806%2Fej.17.1.200
https://cutadapt.readthedocs.org/en/stable/

- **Trimmomatic**:

A flexible read trimming tool for Illumina NGS data
http://www.usadellab.org/cms/?page=trimmomatic

The UniVec Database

UniVec is a database that can be used to quickly identify segments within nucleic acid sequences which may be of vector origin

http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/

ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/

UniVec is designed for use in applications where a scientist will review the hits to weed out the occasional false positive. The sequences included in UniVec are chosen to maximize the detection of contamination with the understanding that a few false positive hits are acceptable.

UniVec_Core is designed for use in applications where the hits will be automatically processed without any human review. The sequences in UniVec_Core are a subset of those from the full UniVec database chosen to minimize the number of false positive hits.

Written in **FASTA** format, can be manipulated or a subdatabase extracted
Ex: Univec database of only Illumina adapters

Adapter cleaning

- **Univec** database (only illumina adapters)
- UCSC **blat** (vecscreen uses too much RAM)

Fast heuristic sequence search tool (searches for similar sequences)
http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/blat/
http://hgdownload.cse.ucsc.edu/admin/jksrc.zip
https://genome.ucsc.edu/FAQ/FAQblat.html
http://www.ncbi.nlm.nih.gov/pubmed/11932250

- **Skewer**

A fast and accurate adapter trimmer for next-generation sequencing paired-end reads
http://www.ncbi.nlm.nih.gov/pubmed/24925680/
http://sourceforge.net/projects/skewer/

- **Cutadapt**

Removes adapter sequences from high-throughput sequencing reads
http://dx.doi.org/10.14806%2Fej.17.1.200
https://cutadapt.readthedocs.org/en/stable/

- **Trimmomatic**:

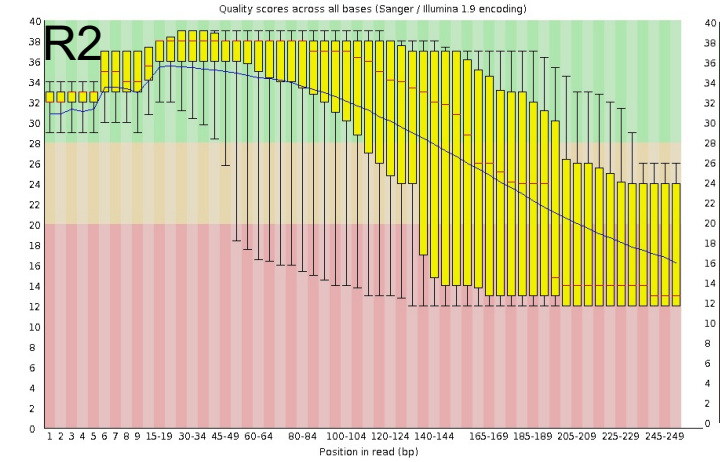
A flexible read trimming tool for Illumina NGS data
http://www.usadellab.org/cms/?page=trimmomatic

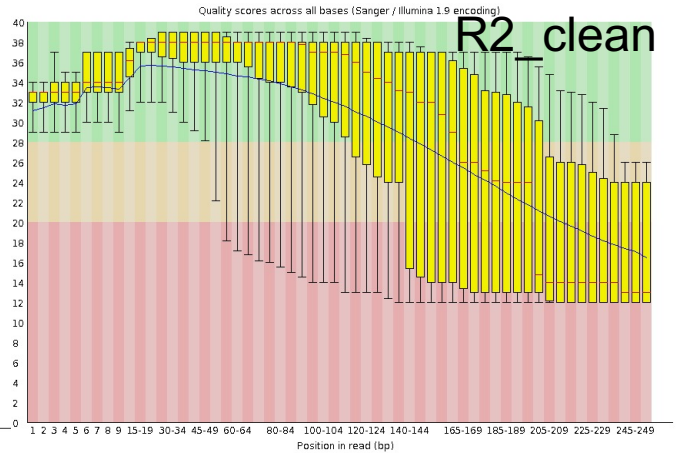
gnl|uv|NGB00735.1:1-47	700
gnl|uv|NGB00736.1:1-47	1578
gnl|uv|NGB00737.1:1-47	88
gnl|uv|NGB00738.1:1-47	692
gnl|uv|NGB00739.1:1-47	38
gnl|uv|NGB00726.1:1-34	62
gnl|uv|NGB00380.1:1-26	432
gnl|uv|NGB00607.1:1563-1661	2
gnl|uv|NGB00031.1:1385-1483	2
gnl|uv|NGB00029.1:270-368	1
gnl|uv|NGB00030.1:1-50	2
gnl|uv|NGB00412.1:3842-3940	1
gnl|uv|NGB00029.1:434-543	1

gnl|uv|NGB00745.1:1-47	52
gnl|uv|NGB00740.1:1-47	31
gnl|uv|NGB00741.1:1-47	779
gnl|uv|NGB00375.1:1-43	19
gnl|uv|NGB00742.1:1-47	12
gnl|uv|NGB00743.1:1-47	29
gnl|uv|NGB00744.1:1-47	29
gnl|uv|NGB00746.1:1-47	653
gnl|uv|NGB00365.1:1-43	19
gnl|uv|NGB00370.1:1-43	18
gnl|uv|NGB00031.1:488-586	1
gnl|uv|NGB00725.1:1-33	3
gnl|uv|NGB00123.1:498-592	1

PG157 S2 R1
Adapters found
Identity > 80%

```
M01269:69:000000000-AD379:1:2114:3542:16836      ACCCTGGCACGCGTGCCAGCGGGGGCTTTCCTCGCCGCCGCCACGCTGTCGCCGGATGGG
gnl|uv|NGB00735.1:1-47                           ------------------------------------------------------------

M01269:69:000000000-AD379:1:2114:3542:16836      CAGCTGGGCACCGGTGCGCTTGCCGCCGGTGCCGGTGTGGCACTGGCCAGTCATGGCCTG
gnl|uv|NGB00735.1:1-47                           ------------------------------------------------------------

M01269:69:000000000-AD379:1:2114:3542:16836      AAGGCCGGCACCCGCGCCCTGCTCAATACCTCGCCCGAGCCCACGAGACTAAGGCGAATC
gnl|uv|NGB00735.1:1-47                           ----------------------------------CCGAGCCCACGAGACTAAGGCGAATC
                                                                                  **************************

M01269:69:000000000-AD379:1:2114:3542:16836      TCGTATGCCGTCTTCTGCTTGAAAAAAAAAAGACAGAAGAGTTGCGATGTGGGGGGGGGG
gnl|uv|NGB00735.1:1-47                           TCGTATGCCGTCTTCTGCTTG--------------------------------------
                                                 ********************

M01269:69:000000000-AD379:1:2114:3542:16836      AGGGTAGAGGG
gnl|uv|NGB00735.1:1-47                           -----------
```

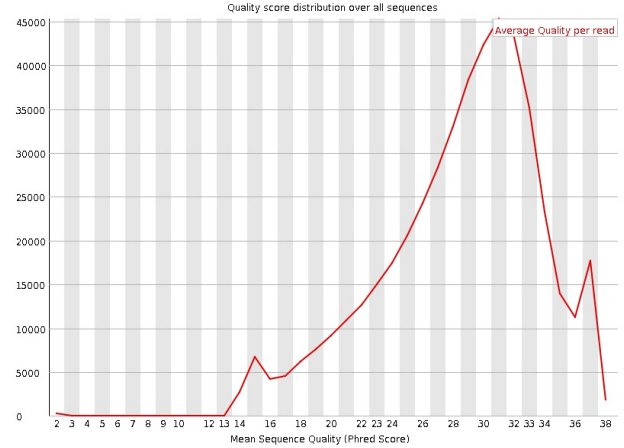


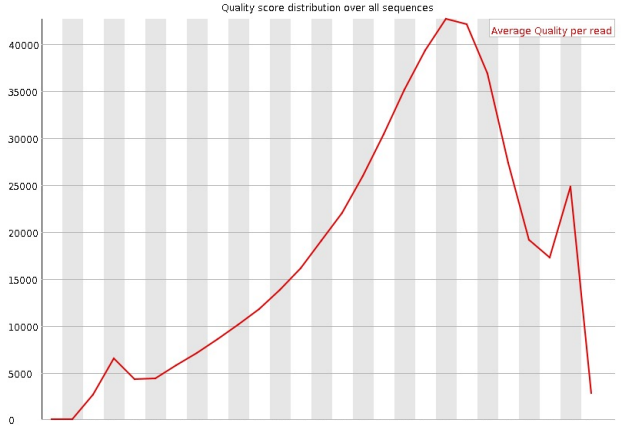


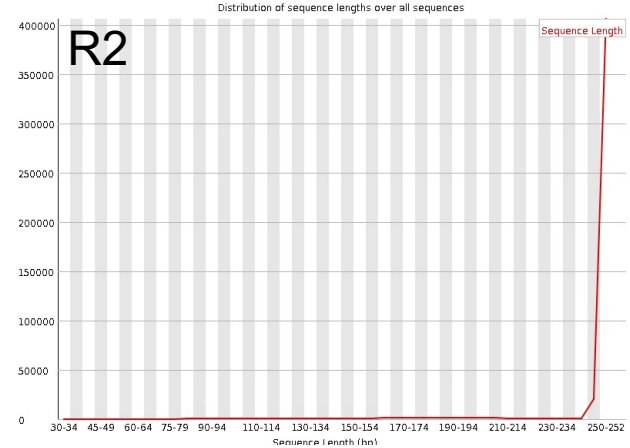
Universitat Autònoma de Barcelona

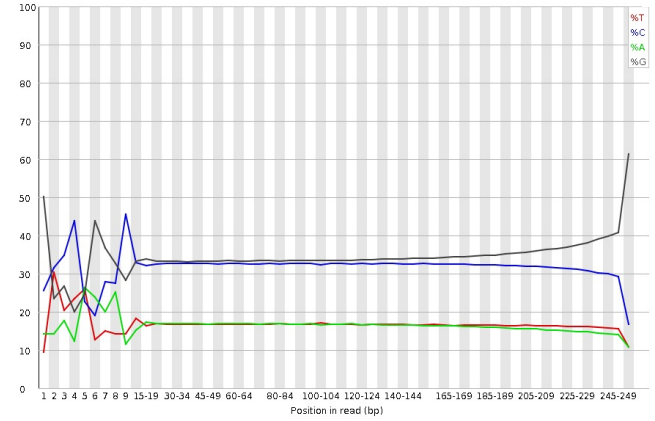
Institut de Biotecnologia i de Biomedicina

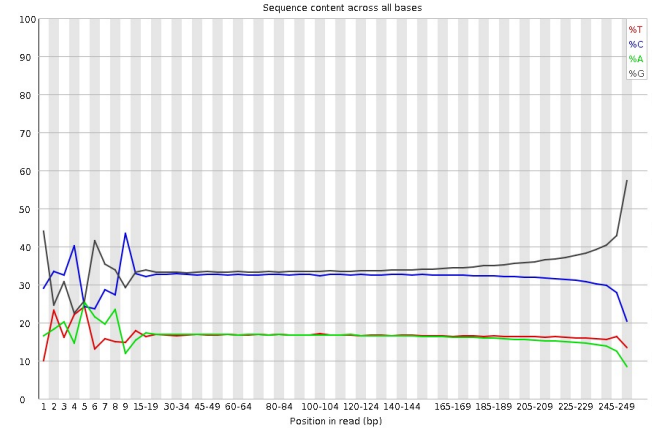
PG157 S2 R2 FastQC before vs after cleaning

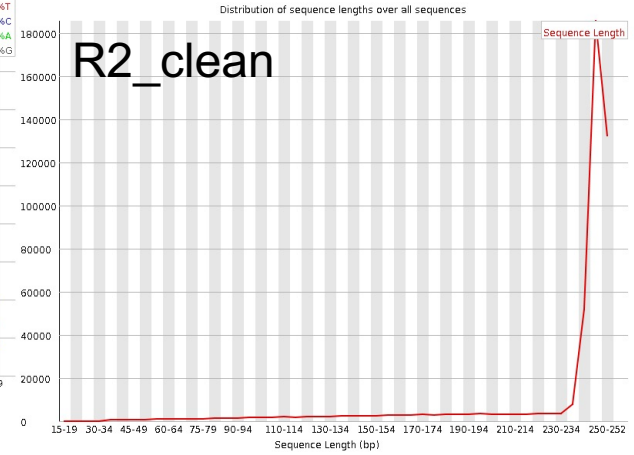




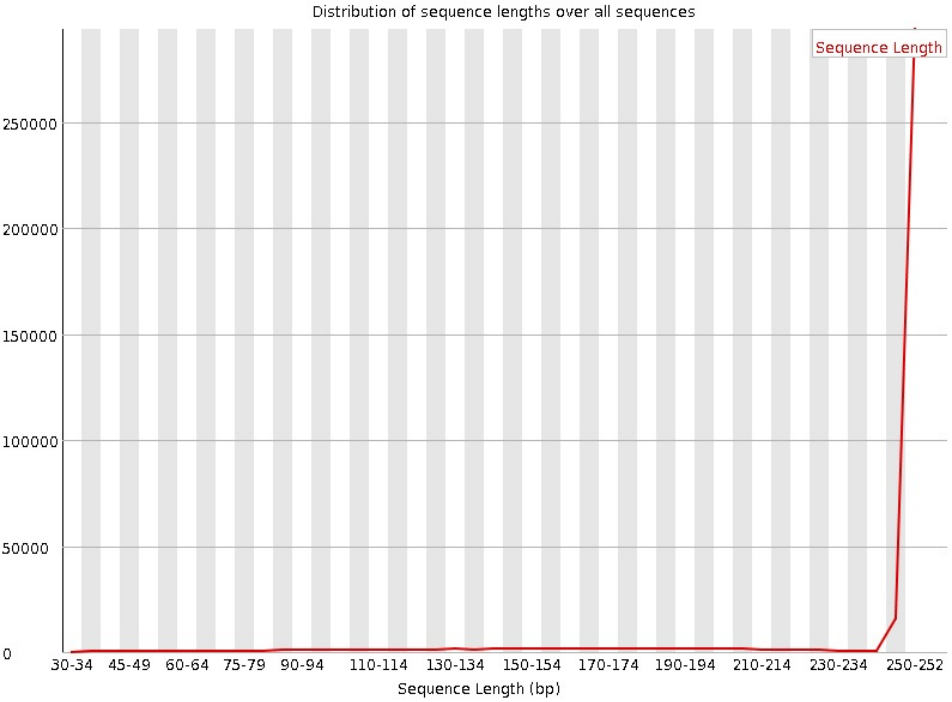


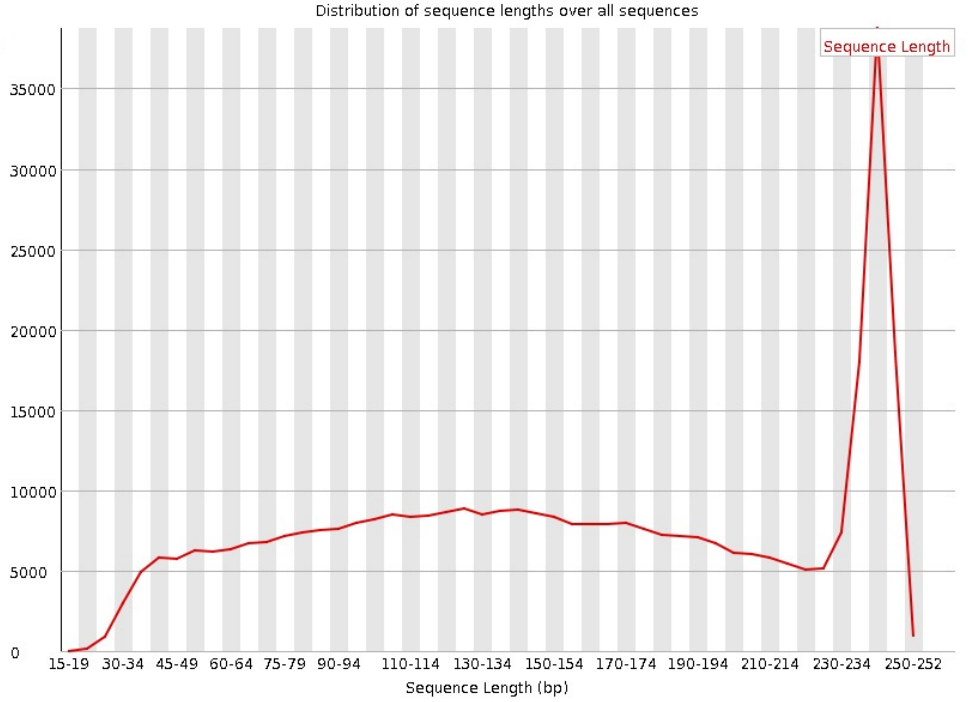




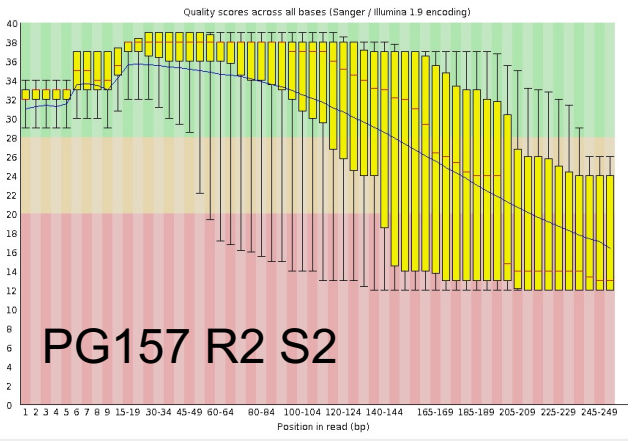


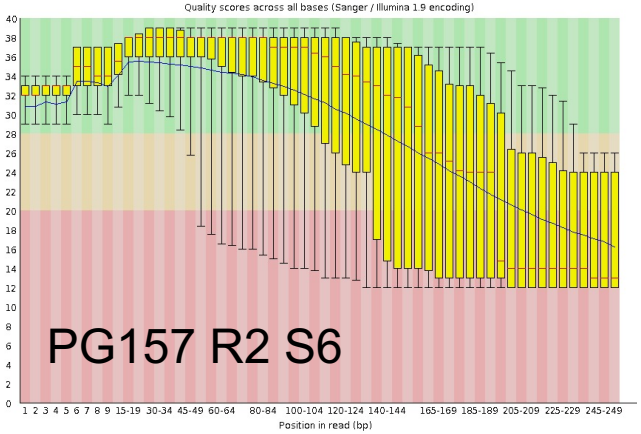
PG157 S6 R2
FastQC
before vs after
cleaning

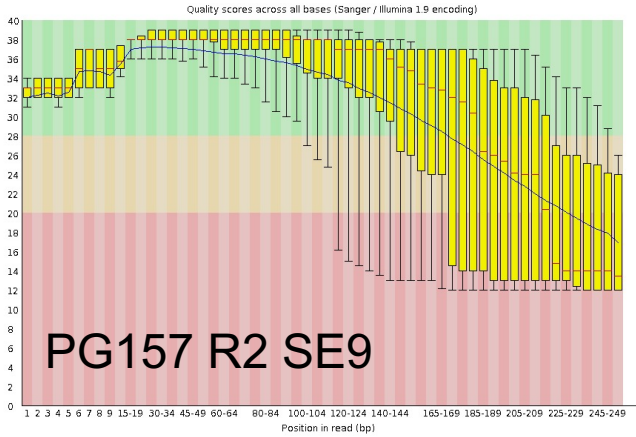




Tail trimming







PG157S2 selected sizes: 140,150,160,170,180
PG157S6 selected sizes: 140,150,160,170,180
PG157SE9 selected sizes: 140,150,160,170,180

Combinations of 1,2 or 3 elements choosen from 3 different lists

```
3*5=     15
5*5*3=   75
5*5*5= 125
```

Total= **215**

Assembly

- **VelvetOptimizer**

Runs multiple instances of **velvet** modifying the Kmer value, outputs the best assembly (only allows 2 samples)

http://bioinformatics.net.au/software.velvetoptimiser.shtml
http://www.ncbi.nlm.nih.gov/pubmed/20836074/
https://www.ebi.ac.uk/~zerbino/velvet/
http://www.ncbi.nlm.nih.gov/pubmed/18349386/

- SPADES

http://bioinf.spbau.ru/spades
http://www.ncbi.nlm.nih.gov/pubmed/24093227

```
${VELOP_EXE} -t ${VELOP_THRDS} -s $KMR_S -e $KMR_E -f "-shortPaired -fastq -separate ../$spl0_pair1 ../$spl0_pair2
-shortPaired2 -separate ../$spl1_pair1 ../$spl1_pair2" -o "-exp_cov auto -scaffolding yes"
```

```
$SPADES_EXE --careful -m $SPA_MEMLIM -t $SPA_THRDS -k 21,33,55,77 --pe1-1 fixed-0_1.fq --pe1-2 fixed-0_2.fq --pe2-1
fixed-1_1.fq --pe2-2 fixed-1_2.fq --pe3-1 fixed-2_1.fq --pe3-2 fixed-2_2.fq -o SPA_OUT/
```

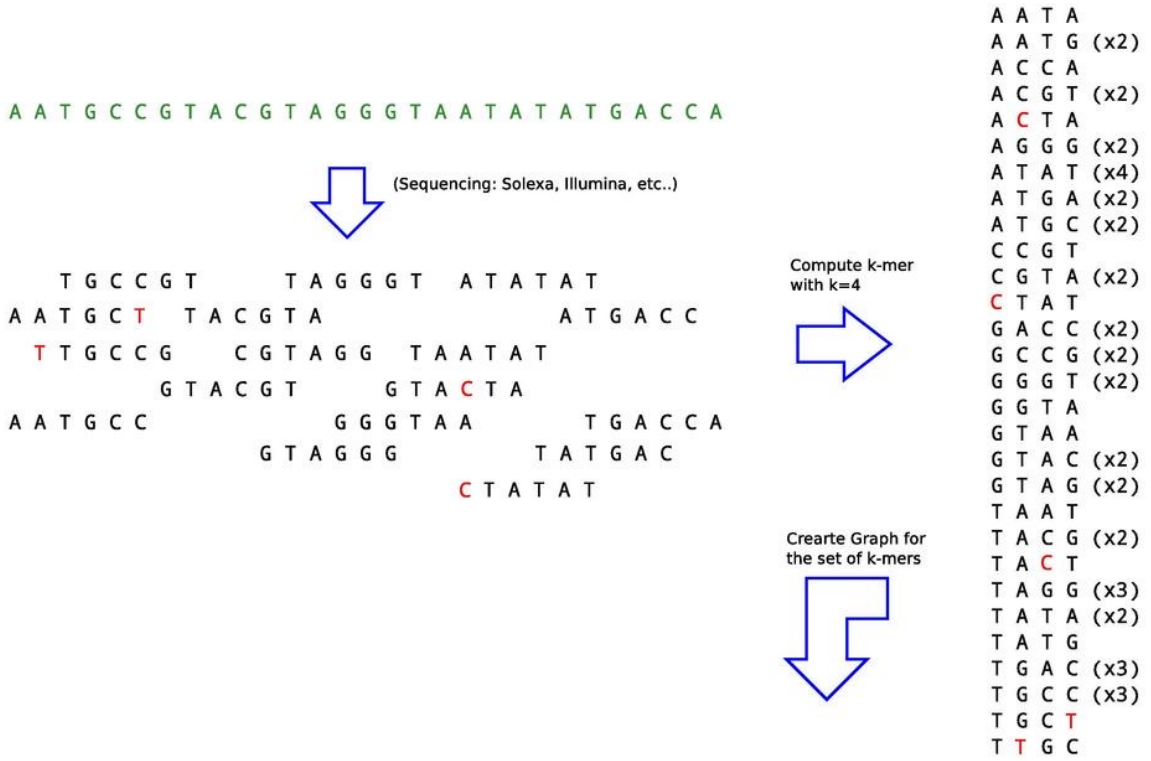
Run the assembly

Perl script to prepare all different runs and submit them to the computer cluster

IBB compuer cluster *Celler* to distribute all calculations and run them simultaneously

```
#!/bin/env perl
use Bio::SeqIO

$SAMPLES{PG157S2}{fastq}[0]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157S2_
$SAMPLES{PG157S2}{fastq}[1]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157S2_
@{$SAMPLES{PG157S2}{trim_sizes}}=(140,150,160,170,180);
$SAMPLES{PG157S6}{fastq}[0]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157S6f
$SAMPLES{PG157S6}{fastq}[1]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157S6f
@{$SAMPLES{PG157S6}{trim_sizes}}=(140,150,160,170,180);
$SAMPLES{PG157ES9}{fastq}[0]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157ES
$SAMPLES{PG157ES9}{fastq}[1]="/share/data0/txino/DNA/STENO/PG157/CLEANED/PG157ES
@{$SAMPLES{PG157ES9}{trim_sizes}}=(140,150,160,170,180);

$DIRS{DATA}="/share/data0/txino/DNA/STENO/PG157";
$DIRS{WORKDIR}="/state/partition1/txino";
$DIRS{TRIM}="$DIRS{DATA}/TRIM";

$SOFT{exe}{BWA}="/share/data0/txino/BIOSOFT/BWA/bwa-0.7.7/bwa";
$SOFT{exe}{samtools}="/share/data0/txino/BIOSOFT/SAMTOOLS/samtools-1.1/bin/samto
$SOFT{exe}{VelOpt}="VelvetOptimiser.pl";
$SOFT{opt}{VelOpt}{threads}=1;
$SOFT{opt}{VelOpt}{kmer_start}=49;
$SOFT{opt}{VelOpt}{kmer_end}=129;
$SOFT{exe}{fastqc}="/share/data0/txino/BIOSOFT/FASTQC/FastQC_0112/fastqc";
$SOFT{exe}{fastqc_old}="/share/data0/txino/BIOSOFT/FASTQC/FastQC_0101/fastqc";
$SOFT{exe}{alex_trim}="/share/data0/txino/BIOSOFT/TRIM/Print_N_Bases.py";
$SOFT{exe}{bbmap_repair}="/share/data0/txino/BIOSOFT/BBMAP/bbmap/repair.sh";
$SOFT{exe}{spades}="/share/data0/txino/BIOSOFT/SPADES/SPAdes-3.5.0-Linux/bin/spa
$SOFT{opt}{spades}{spades_threads}="4";
$SOFT{opt}{spades}{spades_mem_limit}="8";

$SGE{PE}="orte";
$SGE{PE_num}=8;
$SGE{queue}="all.q\@\@notold";



################################################################################
################################################################################

chkargs();

my $hsh_trim_files=&run_sge_trim();

&run_sge_velOp_samples($hsh_trim_files);

print "\nSCRIPT ENDDED, have a nice day!\n";
```

Velvet assembly algorithm: de Bruijn graph



Evaluation

- **Contigs**: Fragments of overlaping reads that can be constructed, the lower the better (if we have 1 chromosome we would like to find only 1 contig, so we have completed the whole genome)

- **n50**: This number indicates that half of the bases of the assembled genome are in contings of this size or longer. It tells us how long are our contigs, the longer the better.

- **Total bases in contigs**: we know, more or less, the size of our organism's genome. The closer to this size, the better. (Stenotrophomonas maltophilia ≈ 4.5 – 4.8 Mbases)

- **Longest conting**

(- N count): number of unidentified residues in the assembly. Sequenced as "N" or introduced during the scaffolding process. Not used for evaluation but has to be taken into account. ("N" residues are not allowed by NCBI)

Used_samples	contigs	n50	bases in contigs	longest contig	
PG157S2_170-PG157ES9_150-scaf	237	43794	4769914	146238	VelvetOptimiser
PG157S2_170-PG157ES9_150-noscf	245	41659	4769055	146238	
PG157S2_160-PG157ES9_150-scaf	263	37324	4772324	142476	
PG157S2_180-PG157ES9_140-noscf	247	37318	4769264	146238	
PG157S6_150-PG157S2_180-scaf	309	36679	4779216	104459	
PG157S2_180-PG157ES9_160-noscf	251	35077	4771950	138222	
PG157S2_160-PG157ES9_170-noscf	257	35071	4771125	138223	
PG157S2_150-PG157ES9_170-noscf	276	34948	4772847	138223	
PG157S6_180-PG157S2_140-PG157ES9_140-SPsca	177	143740	4915964	459235	Spades
PG157S6_150-PG157S2_150-PG157ES9_180-SPsca	183	143740	4835278	459235	
PG157S6_180-PG157S2_140-PG157ES9_140-SPcon	184	143740	4915069	293555	
PG157S6_180-PG157S2_160-PG157ES9_140-SPsca	202	143740	4865251	459236	
PG157S6_150-PG157S2_160-PG157ES9_180-SPsca	203	143740	4863291	459235	
PG157S6_180-PG157S2_160-PG157ES9_140-SPcon	209	143740	4864367	293555	
PG157S6_160-PG157S2_180-PG157ES9_150-SPsca	221	143740	4869193	459235	
PG157S6_180-PG157S2_170-PG157ES9_140-SPsca	223	143740	4932197	459235	

Refinement: PAGIT

Contig reorder and extension. “N” removal

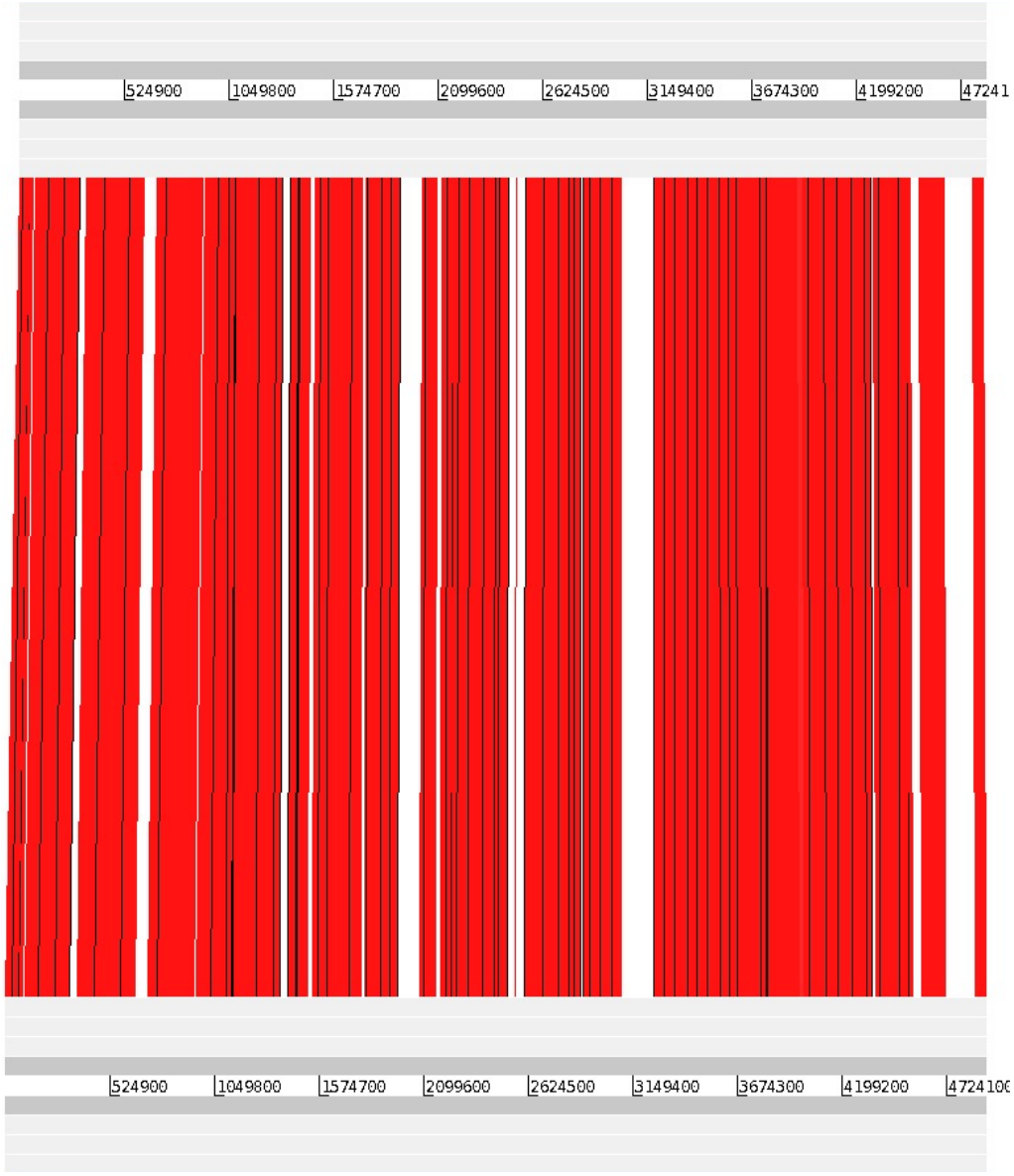
http://www.sanger.ac.uk/science/tools/pagit
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3648784/

- ABACAS is able to contiguate contigs from a de novo assembly against a closely related reference.
http://www.ncbi.nlm.nih.gov/pubmed/19497936

- IMAGE, an iterative approach for closing gaps in assembled genomes using mate pair information. It is able to close gaps left open by the assembler in a draft genome, even when using the same data sets as used by the original assembler.
http://www.ncbi.nlm.nih.gov/pubmed/20388197

PAGIT: ABACAS

```
perl $PAGIT_HOME/ABACAS/abacas.pl -r reference.fna -q contigs.fa -p nucmer -m -b -c -o PG157vo_abcs
```

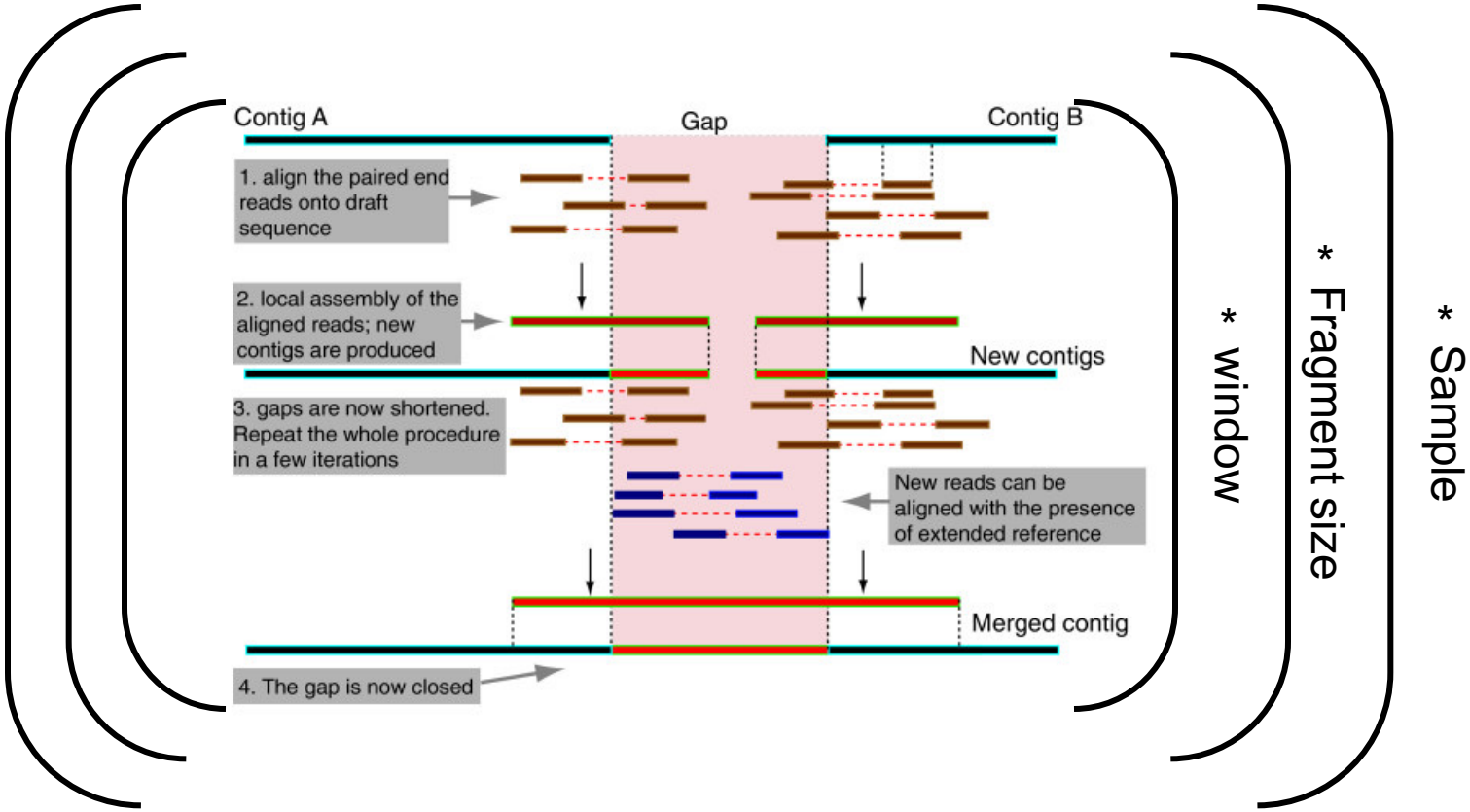


Contigs not present in reference

sum = 775651, n = 33,
ave = 23504.58, largest = 146238
N50 = 77843, n = 4
N60 = 70913, n = 5
N70 = 59194, n = 6
N80 = 36729, n = 8
N90 = 21590, n = 11
N100 = 178, n = 33
N_count = 66

PAGIT: IMAGE

```
perl $PAGIT_HOME/IMAGE/image.pl -scaffolds contig.namefix.fa -prefix paired -iteration 1 -all_iteration 3 -
dir_prefix ite -kmer 69

restartIMAGE.pl ite3 65 3 partitioned
...
restartIMAGE.pl ite117 33 3 partitioned
```



We have several samples and several sizes for sample, so we run it several times reiterating for each sample and for each size (perl script)

before PAGIT (VelOpt run)

sum = 4769914, n = 237, ave = 20126.22, largest = 146238
N50 = 43794, n = 35
N60 = 35063, n = 48
N70 = 26927, n = 63
N80 = 20827, n = 84
N90 = 12214, n = 114
N100 = 178, n = 237
N_count = 150

After PAGIT
sum = 4762953, n = 201, ave = 23696.28, largest = 146225
N50 = 42801, n = 37
N60 = 32485, n = 49
N70 = 26801, n = 65
N80 = 19119, n = 86
N90 = 12214, n = 117
N100 = 565, n = 201
N_count = 0

MIX

We have now two different assemblies from two different methods, both have data the other one does not. We need to combine them.

Mix algorithm takes two assemblies and generates another one that mixes them in order to extend the length of resulting contigs. It builds an assembly graph in which all of the contigs are vertices and edges represent the best possible alignments between two contigs that have the potential of being used as basis for contig extension.The resulting output assembly corresponds to a certain path in this assembly graph.
https://github.com/cbib/MIX
http://www.ncbi.nlm.nih.gov/pubmed/24564706

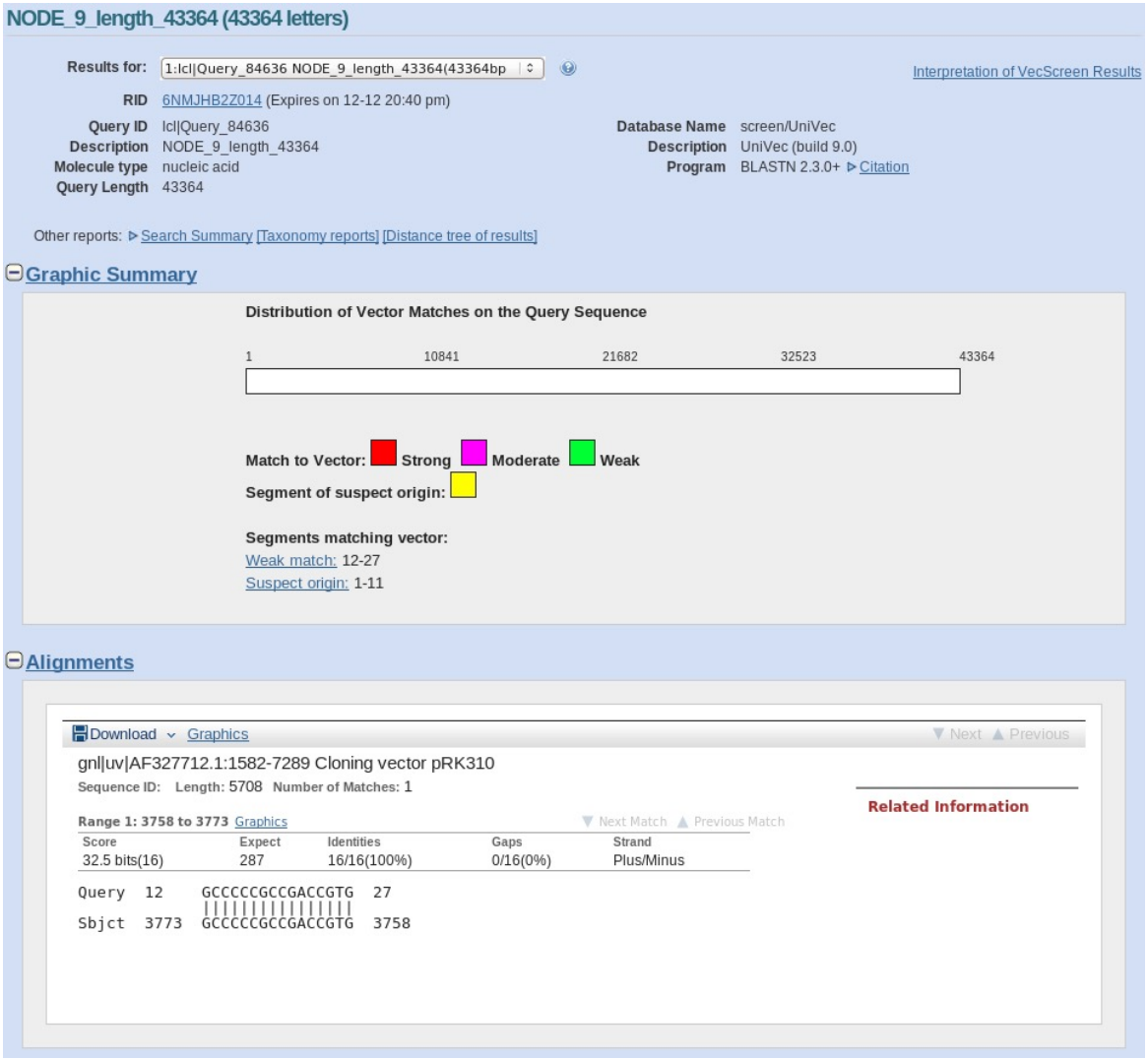
```
$MIX_BIN_DIR/preprocessing.py VOcontigs.fa SPAcontigs.fa -o all_contigs.fa
$MUMER_DIR/nucmer -prefix=alignments all_contigs.fa all_contigs.fa
$MUMER_DIR/show-coords -rcl alignments.delta > alignments.coords
mkdir MIX_OUT
$MIX_BIN_DIR/Mix.py -a alignments.coords -o MIX_OUT -c all_contigs.fa -A 500 -C 0 -g
```

VelOp	Spades	MIX (after vecsreen)
sum = 4762953, n = 201,	sum = 4894928, n = 77,	sum = 4949420, n = 76,
ave = 23696.28, largest = 146225	ave = 63570.49, largest = 459235	ave = 65123.95, largest = 459235
N50 = 42801, n = 37	N50 = 144111, n = 12	N50 = 144111, n = 12
N60 = 32485, n = 49	N60 = 109588, n = 16	N60 = 109588, n = 16
N70 = 26801, n = 65	N70 = 82547, n = 21	N70 = 82547, n = 21
N80 = 19119, n = 86	N80 = 48535, n = 29	N80 = 48535, n = 29
N90 = 12214, n = 117	N90 = 31613, n = 41	N90 = 33496, n = 40
N100 = 565, n = 201	N100 = 609, n = 77	N100 = 609, n = 76
N_count = 0	N_count = 0	N_count = 0

VECSCREEN Final contaminant check

http://www.ncbi.nlm.nih.gov/tools/vecscreen/

```
$VSCREEN_EXE -db $UNIVEC_DB -query $1 -out $1.vscrn0 -outfmt 0 -text_output
```

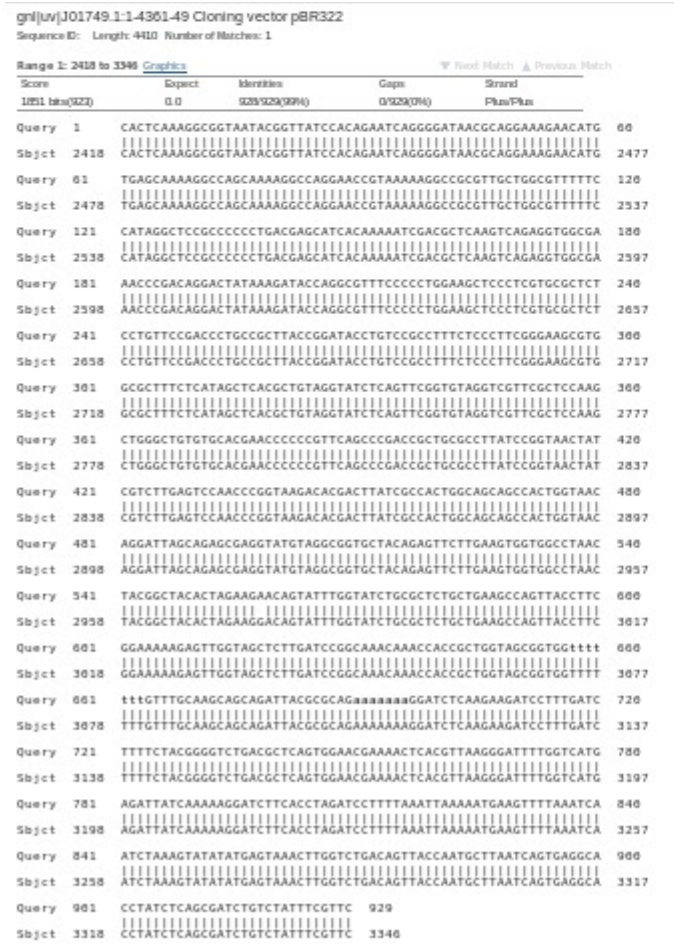


gnl|uv|J01749.1:1-4361-49 Cloning vector pBR322

Sequence ID: **Length:** 4410 **Number of Matches:** 1

Range 1: 2418 to 3346 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1851 bits(923)	0.0	928/929(99%)	0/929(0%)	Plus/Plus



Results for: 2:lcl|Query_84637 NODE_70_length_929(929bp)

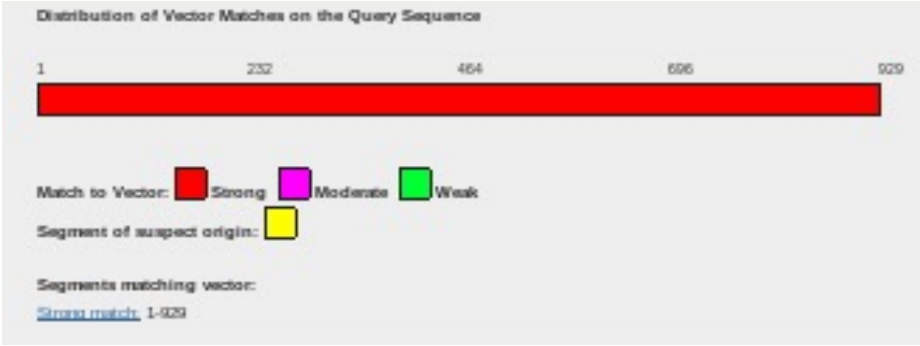
RID 6NMJHB2Z014 (Expires on 12-12 20:40 pm)

Query ID lcl|Query_84637

Description NODE_70_length_929

Molecule type nucleic acid

Query Length 929



COVERAGE & Final assembly check

Reference alignment: Final_assembly vs shortest trimmed reads from each sample

- BWA

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome

http://bio-bwa.sourceforge.net/
https://github.com/lh3/bwa
http://www.ncbi.nlm.nih.gov/pubmed/20080505
http://arxiv.org/abs/1303.3997

- Samtools

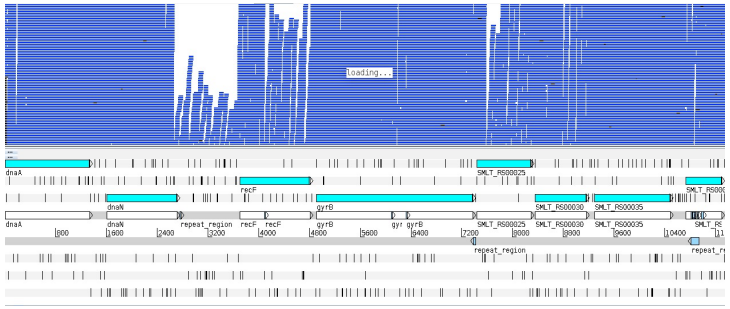
Samtools is a suite of programs for interacting with high-throughput sequencing data.

http://samtools.sourceforge.net/
http://www.htslib.org/
https://www.ncbi.nlm.nih.gov/pubmed/19505943

```
# Create index for the reference (assembly)
$BWA_EXE." index -p ".$bwaidx_head." -a is $$hshfl_asmblfasta{name}

# align (one run for each sample)
$BWA_EXE." mem -t ".$N_THREADS." ".$$hsh_bwaidx{idxpathname}." ".$fq_filename[0]." ".$fq_filename[1]."
>".$sam_file

# convert SAM to BAM
$SAMTOOLS_EXE." view -S ".$sam_file." -b >".$bam_file

# sort by name
$SAMTOOLS_EXE." sort -n ".$bam_file." ".$namesort_bam_file

#fix mate pairs
$SAMTOOLS_EXE." fixmate ".$namesort_bam_file.".bam ".$fix_bam_file.".bam"

# sort by coordinates
$SAMTOOLS_EXE." sort ".$fix_bam_file.".bam ".$sort_bam_file

# remove duplicates
$SAMTOOLS_EXE." rmdup ".$sort_bam_file.".bam ".$rmdp_bam_file.".bam"

#variant calling (sort of)
$SAMTOOLS_EXE." mpileup -C50 -f ".$fasta_file." ".$$hsh_all_bams{rmdp}." >".$rmdp_mpileup_file
```

PG157S2

976445 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
960376 + 0 mapped (**98.35**%:-nan%)
976445 + 0 paired in sequencing
488451 + 0 read1
487994 + 0 read2
900557 + 0 properly paired (**92.23**%:-nan%)
919823 + 0 with itself and mate mapped
40553 + 0 singletons (4.15%:-nan%)
5402 + 0 with mate mapped to a different chr
4477 + 0 with mate mapped to a different chr (mapQ>=5)

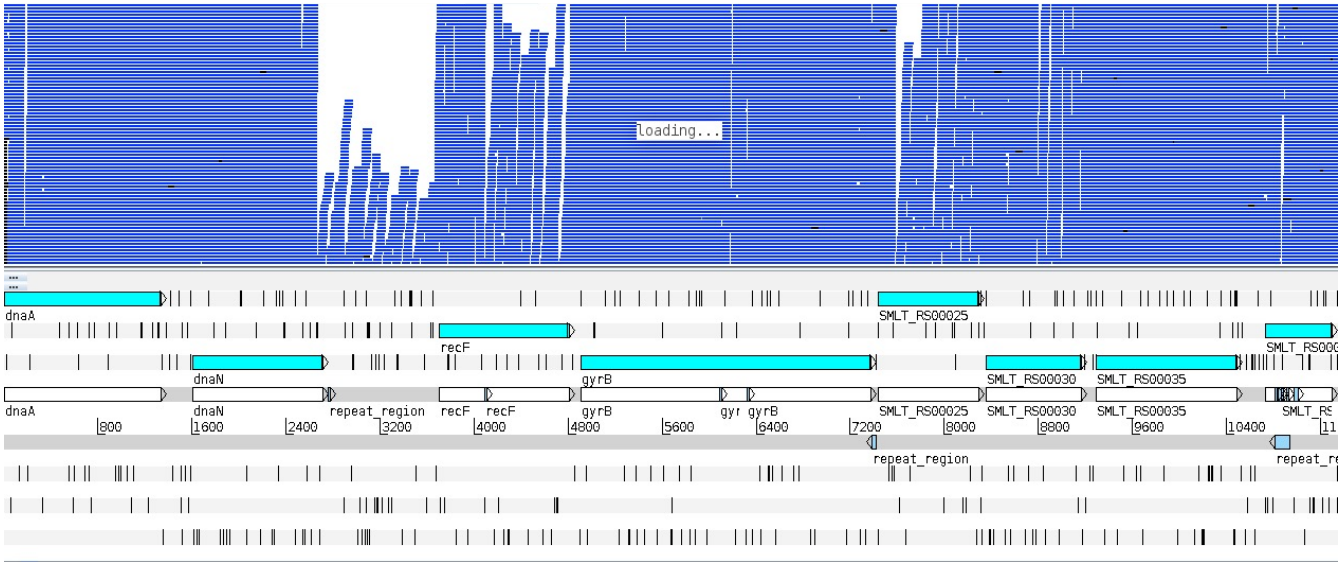


PG157S6

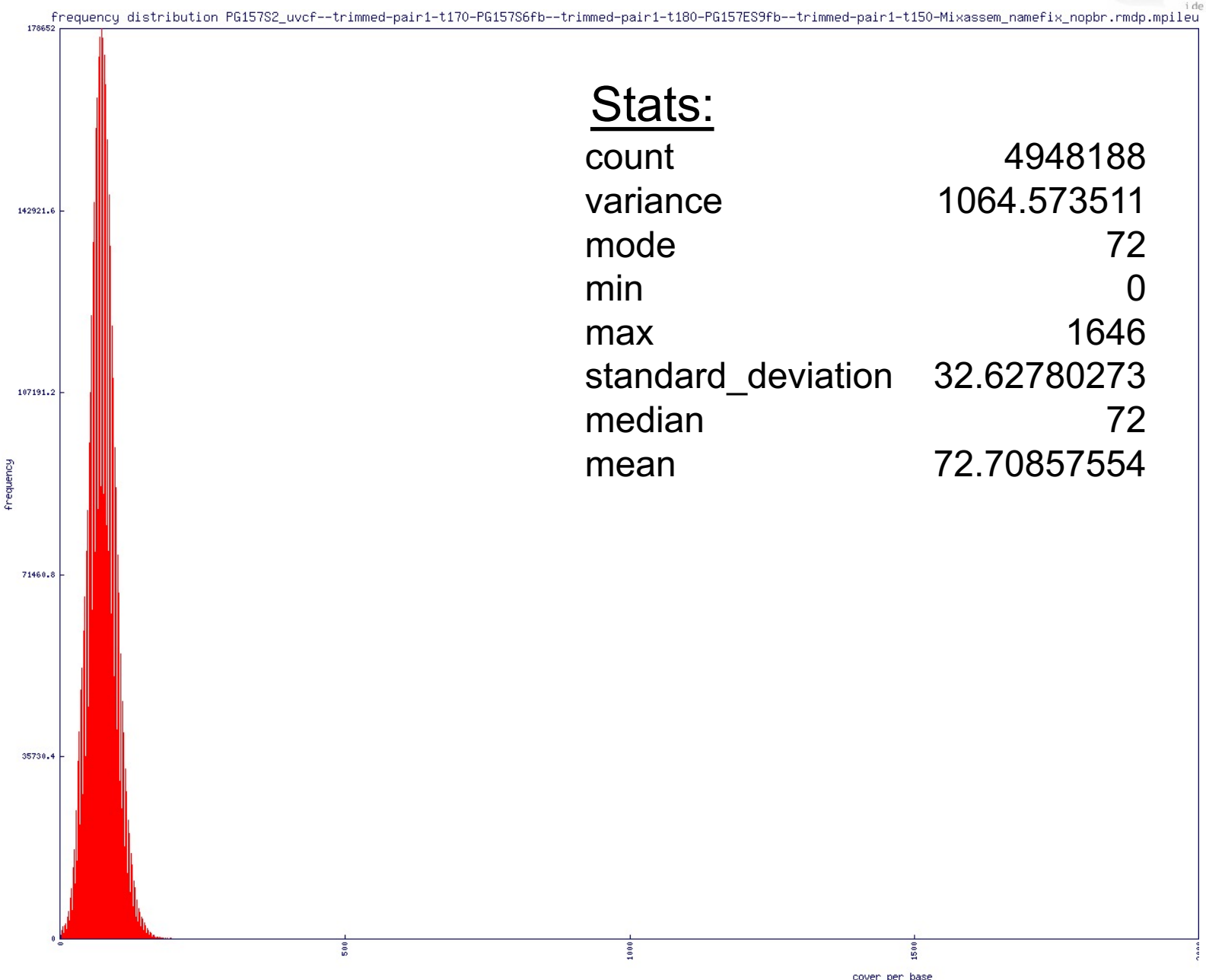
754302 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
749180 + 0 mapped (**99.32**%:-nan%)
754302 + 0 paired in sequencing
378868 + 0 read1
375434 + 0 read2
717881 + 0 properly paired (**95.17**%:-nan%)
730183 + 0 with itself and mate mapped
18997 + 0 singletons (2.52%:-nan%)
4996 + 0 with mate mapped to a different chr
4231 + 0 with mate mapped to a different chr (mapQ>=5)

PG157ES9

932317 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
929196 + 0 mapped (**99.67**%:-nan%)
932317 + 0 paired in sequencing
466234 + 0 read1
466083 + 0 read2
891184 + 0 properly paired (**95.59**%:-nan%)
912550 + 0 with itself and mate mapped
16646 + 0 singletons (1.79%:-nan%)
2742 + 0 with mate mapped to a different chr
2146 + 0 with mate mapped to a different chr (mapQ>=5)



NODE_1_length_208705	153	C	9	.......,^S,	HHHGGFFDE	6	...,..	HHHFFA	7	...,...	EHHFHHB
NODE_1_length_208705	154	G	9	........,,	GGFGGCAAF	6	...,..	GGGFCA	7	...,...	GDGGGGC
NODE_1_length_208705	155	C	9	........,,	GGGGGGBBF	6	...,..	GGGFCA	7	...,...	EAGCGGC
NODE_1_length_208705	156	C	9	........,,	GGGGGGB@F	6	...,..	GGGFC?	7	...,...	GAGDGGC
NODE_1_length_208705	157	G	9	........,,	GGGGGGB;F	6	...,..	GGGFG3	7	...,...	GCGCGGB
NODE_1_length_208705	158	C	8	........,	GGGGGGBF	6	...,..	GGGFGA	7	...,...	GFGBGGC
NODE_1_length_208705	159	A	9	........,,	GGCEGGB9F	6	...,..	GGGFG2	7	...,...	GFG?GGC
NODE_1_length_208705	160	C	9	........,,	GGGGGGGF=	6	...,..	GGGFGA	7	...,...	G/G?GGC
NODE_1_length_208705	161	C	9	........,,	GGGGGGG;A	6	...,..	GGGFGD	7	...,...	GBGFGGC
NODE_1_length_208705	162	G	9	........,,	GGGGGGGAF	6	...,..	GGGFGD	7	...,...	GDG?GGD
NODE_1_length_208705	163	C	9	........,,	GGGGGGGAF	6	...,..	GGGFGB	7	...,...	GGGFGGC
NODE_1_length_208705	164	G	9	........,,	GGG?GGGFE	6	...,..	GGGFGG	7	...,...	GCGGGGG
NODE_1_length_208705	165	C	9	........,,	GGG?GGGFE	6	...,..	GGGEGG	7	...,...	GGGHGGG
NODE_1_length_208705	166	T	9	........,,	GGGEGGG=A	6	...,..	GGGFGG	7	...,...	GCGGGGG



Stats:

count	4948188
variance	1064.573511
mode	72
min	0
max	1646
standard_deviation	32.62780273
median	72
mean	72.70857554

Genome Annotation

- **RAST** (Rapid Annotation using Subsystem Technology)
is a fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylogenetic tree.
http://rast.nmpdr.org/
http://www.ncbi.nlm.nih.gov/pubmed/18261238

- **PGAP** (NCBI Prokaryotic Genome Annotation Pipeline)
NCBI has developed an automatic annotation pipeline that combines ab initio gene prediction algorithms with homology based methods
http://www.ncbi.nlm.nih.gov/genome/annotation_prok/
http://www.ncbi.nlm.nih.gov/pubmed/18416670

RAST

Organism Overview for Stenotrophomonas maltophilia PG157mix (6666666.155017)

Genome	Stenotrophomonas maltophilia PG157mix
Domain	Bacteria
Taxonomy	Bacteria; Stenotrophomonas maltophilia PG157mix
Neighbors	View closest neighbors
Size	4,950,349 bp
Number of Contigs (with PEGs)	77
Number of Subsystems	446
Number of Coding Sequences	4688
Number of RNAs	79

For each genome we offer a wide set of information to browse, compare and download.

Browse | Compare | Download | Annotate

Browse through the features of Stenotrophomonas maltophilia PG157mix both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

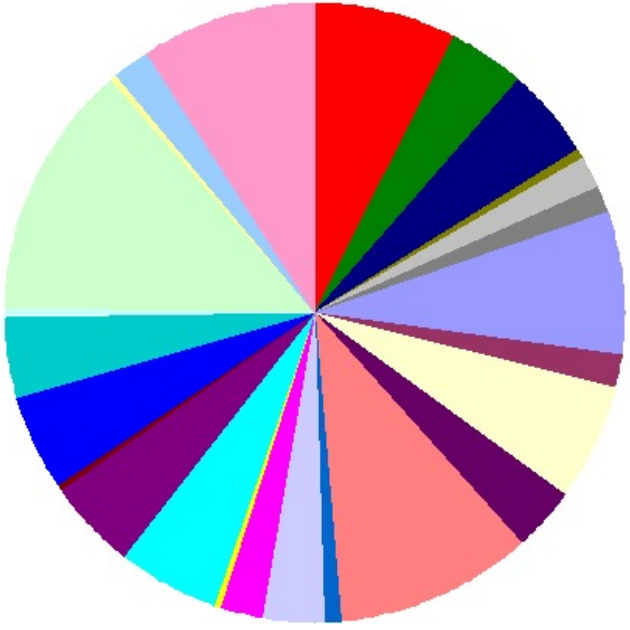
Click here to get to the Genome Browser

Subsystem Information

Subsystem Statistics | Features in Subsystems

Subsystem Coverage

Subsystem Category Distribution



Subsystem Feature Counts

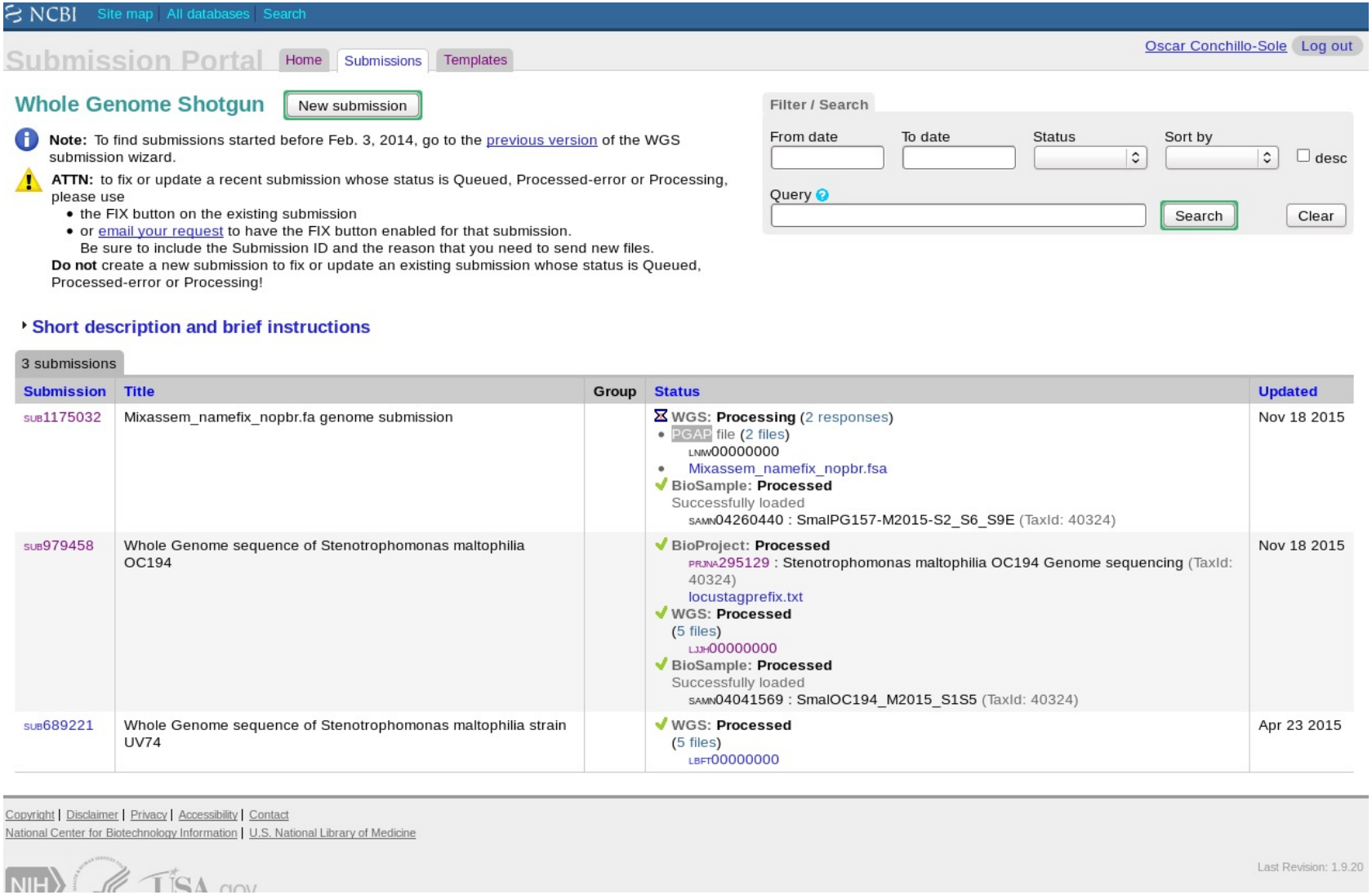
- Cofactors, Vitamins, Prosthetic Groups, Pigments (206)
- Cell Wall and Capsule (104)
- Virulence, Disease and Defense (127)
- Potassium metabolism (13)
- Photosynthesis (0)
- Miscellaneous (47)
- Phages, Prophages, Transposable elements, Plasmids (30)
- Membrane Transport (201)
- Iron acquisition and metabolism (47)
- RNA Metabolism (161)
- Nucleosides and Nucleotides (89)
- Protein Metabolism (276)
- Cell Division and Cell Cycle (26)
- Motility and Chemotaxis (89)
- Regulation and Cell signaling (55)
- Secondary Metabolism (8)
- DNA Metabolism (143)
- Fatty Acids, Lipids, and Isoprenoids (125)
- Nitrogen Metabolism (8)
- Dormancy and Sporulation (2)
- Respiration (134)
- Stress Response (113)
- Metabolism of Aromatic Compounds (16)
- Amino Acids and Derivatives (350)
- Sulfur Metabolism (9)
- Phosphorus Metabolism (51)
- Carbohydrates (241)

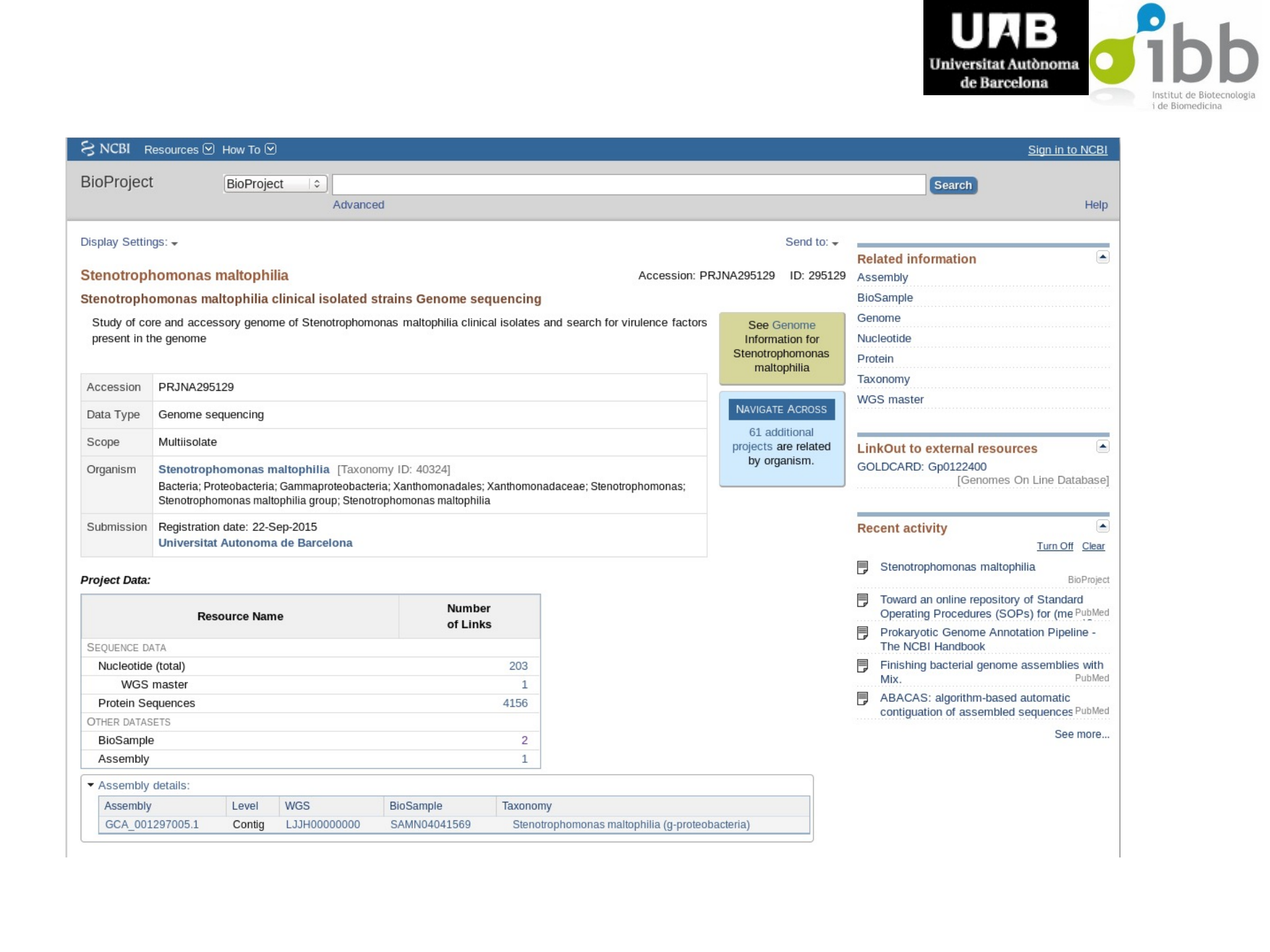
+ Fast & simple
+ Extra tools & info

- Not always agrees with PGAP

NCBI submission

https://submit.ncbi.nlm.nih.gov/subs/wgs/




NCBI Resources How To
Sign in to NCBI
BioProject
BioProject
Search
Advanced
Help
Display Settings:
Send to:
Stenotrophomonas maltophilia
Accession: PRJNA295129 ID: 295129
Stenotrophomonas maltophilia clinical isolated strains Genome sequencing
Study of core and accessory genome of Stenotrophomonas maltophilia clinical isolates and search for virulence factors present in the genome
See Genome Information for Stenotrophomonas maltophilia
NAVIGATE ACROSS
61 additional projects are related by organism.
Accession PRJNA295129
Data Type Genome sequencing
Scope Multiisolate
Organism Stenotrophomonas maltophilia [Taxonomy ID: 40324]
Bacteria; Proteobacteria; Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae; Stenotrophomonas; Stenotrophomonas maltophilia group; Stenotrophomonas maltophilia
Submission Registration date: 22-Sep-2015
Universitat Autonoma de Barcelona
Project Data:
Resource Name
Number of Links
SEQUENCE DATA
Nucleotide (total) 203
WGS master 1
Protein Sequences 4156
OTHER DATASETS
BioSample 2
Assembly 1
Assembly details:
Assembly Level WGS BioSample Taxonomy
GCA_001297005.1 Contig LJJH00000000 SAMN04041569 Stenotrophomonas maltophilia (g-proteobacteria)
Related information
Assembly
BioSample
Genome
Nucleotide
Protein
Taxonomy
WGS master
LinkOut to external resources
GOLDCARD: Gp0122400
[Genomes On Line Database]
Recent activity
Turn Off Clear
Stenotrophomonas maltophilia
BioProject
Toward an online repository of Standard Operating Procedures (SOPs) for (me
PubMed
Prokaryotic Genome Annotation Pipeline - The NCBI Handbook
Finishing bacterial genome assemblies with Mix.
PubMed
ABACAS: algorithm-based automatic contiguation of assembled sequences
PubMed
See more...

Full Send to:

Pathogen: clinical or host-associated sample from Stenotrophomonas maltophilia

Identifiers BioSample: SAMN04260440; Sample name: SmalPG157-M2015-S2_S6_S9E

Organism Stenotrophomonas maltophilia
cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae; Stenotrophomonas; Stenotrophomonas maltophilia group

Package Pathogen: clinical or host-associated; version 1.0

Attributes

strain	PG157
collected by	Golnik Center for Respiratory and Allergic Diseases
collection date	2011-11-14
geographic location	Slovenia:Golnik
host	Homo sapiens
host disease	unknown
isolation source	perineum
latitude and longitude	46.33 N 14.330000000000041 E
host age	73
host subject id	15075

BioProject PRJNA295129 Stenotrophomonas maltophilia
Retrieve all samples from this project

Submission Universitat Autonoma de Barcelona, Oscar Conchillo-Sole; 2015-11-12

Accession: SAMN04260440 ID: 4260440
BioProject

http://www.ncbi.nlm.nih.gov/genome

Stenotrophomonas maltophilia

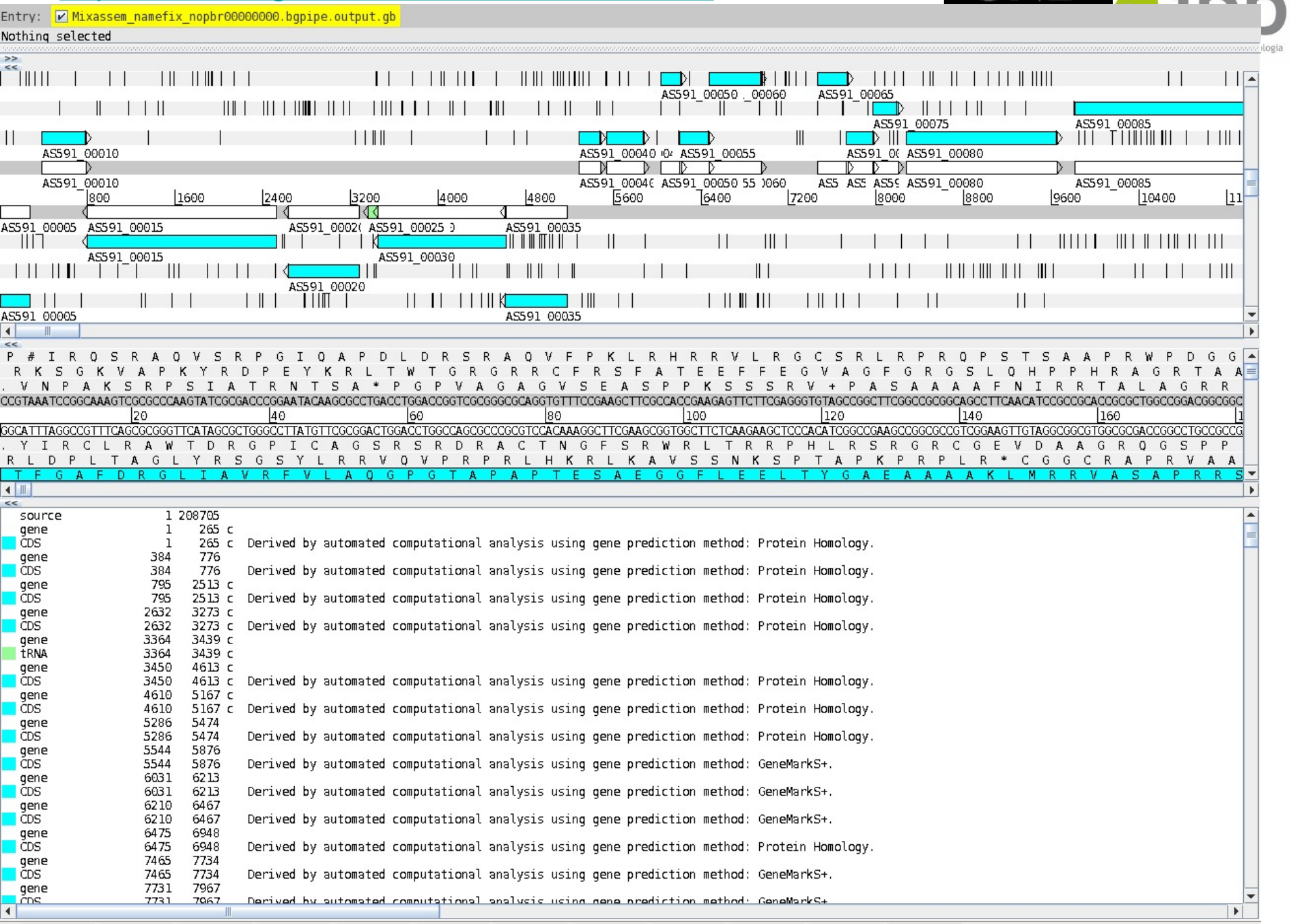
Partial: All **Anomalous:** All **Levels:** ☐ All ☑ ●Complete [7] ☑ ◕Chromosome [2] ☐ ◑Scaffold [52] ☑ ◔Contig [46]

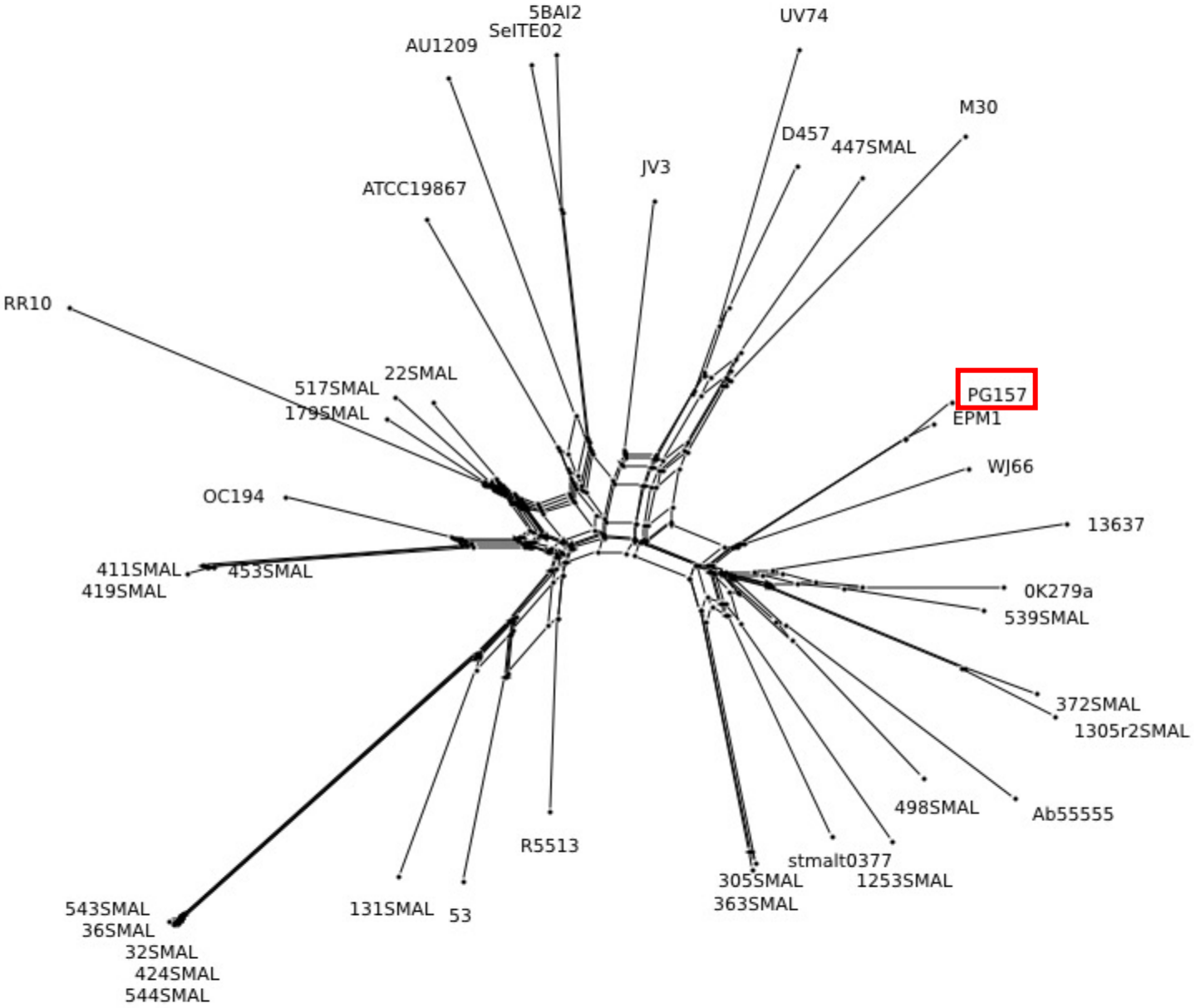
Items 1 - 55 of 55 << First < Prev

Organism/Name	Strain	CladeID	BioSample	BioProject	Assembly	Level	Size (Mb)	GC %	Replicons	WGS	Scaffolds	Gene	Protein
Stenotrophomonas maltophilia K279a	K279a	19485	SAMEA1705934	PRJNA30351	GCA_000072485.1	●	4.85113	66.30	chromosome:NC_010943.1/AM743169.1			4458	4328
Stenotrophomonas maltophilia 5BA-I-2	5BA-I-2	35021	SAMN02641498	PRJNA229945	GCA_000543365.1	◔	4.57397	66.30	-	AZAE01	4	4112	4011
Stenotrophomonas maltophilia R551-3	R551-3	19485	SAMN00623065	PRJNA17107	GCA_000020665.1	●	4.57397	66.30	chromosome:NC_011071.1/CP001111.1	-	-	4112	4011
Stenotrophomonas maltophilia JV3	JV3	19485	SAMN02261377	PRJNA53943	GCA_000223885.1	●	4.54448	66.90	chromosome:NC_015847.1/CP002986.1	-	-	4128	4022
Stenotrophomonas maltophilia D457	D457	19485	SAMEA2272378	PRJEA89665	GCA_000284595.1	●	4.[illegible]	[illegible]	chromosome:[illegible]671.1/HE798556.1	-	-	4381	4220
Stenotrophomonas maltophilia	ISMMS3	19485	SAMN03389650	PRJNA277366	GCA_001274595.1	●	4.804	66.70	chromosome:NZ_CP011045.1/CP011045.1	-	-	4335	4217
Stenotrophomonas maltophilia	ISMMS2	19485	SAMN03389647	PRJNA277366	GCA_001274655.1	●	4.[illegible]	[illegible]	chromosome:NZ_CP011305.1/CP011305.1	-	-	[illegible]	[illegible]
Stenotrophomonas maltophilia	ISMMS2R	19485	SAMN03389649	PRJNA277366	GCA_001274675.1	●	4.50972	66.40	chromosome:NZ_CP011306.1/CP011306.1	-	-	4076	3964
Stenotrophomonas maltophilia	13637	19485	SAMN02874005	PRJNA244350	GCA_000742995.1	◕	4.[illegible]	[illegible]	chromosome:NZ_CP[illegible]8.1/CP[illegible]838.1	-	-	[illegible]	[illegible]
Stenotrophomonas maltophilia EPM1	EPM1	19485	SAMN02471395	PRJNA165731	GCA_000344215.1	◕	4.78777	66.40	chromosome:NZ_CM001824.1/CM001824.1	AMXM01	1	4382	4258
Stenotrophomonas maltophilia RR-10	RR-10	19485	SAMN02471024	PRJNA74289	GCA_000237025.2	◔	[illegible]	[illegible]	-	AGRB01	158	4311	4055
Stenotrophomonas maltophilia S028	S028	21195	SAMN02469568	PRJNA173046	GCA_000295735.2	◔	3.75475	67.10	-	ALYK02	297	3466	3149
Stenotrophomonas maltophilia PML168	PML168	21195	SAMEA2272452	PRJEB46	GCA_000308335.1	◔	[illegible]	[illegible]	-	CAJH01	97	3977	3838
Stenotrophomonas maltophilia AU12-09	AU12-09	19485	SAMN02469852	PRJNA174752	GCA_000346445.1	◔	4.5473	66.40	-	APIT01	125	4209	4025
Stenotrophomonas maltophilia RA8	RA8	19485	SAMEA2272195	PRJEA77887	GCA_000355725.1	◔	4.85145	65.70	-	CALM01	1363	4571	4127
Stenotrophomonas maltophilia SKK35	SKK35	-	SAMEA2272795	PRJEA77885	GCA_000355745.1	◔	4.42031	66.80	-	CALN01	1418	4363	3109
Stenotrophomonas maltophilia stmalt0435	stmalt0435	19485	SAMEA2272536	PRJEB3963	GCA_000455625.1	◔	4.71082	66.40	-	CBQU01	219	4397	4154
Stenotrophomonas maltophilia MF89	MF89	19485	SAMN02471816	PRJNA202921	GCA_000455685.1	◔	4.64903	66.20	-	ATAP01	209	4204	4091
Stenotrophomonas maltophilia stmalt0377	stmalt0377	19485	SAMEA3138820	PRJEB3410	GCA_000499565.1	◔	4.62084	66.40	-	CBQT01	120	4283	4104
Stenotrophomonas maltophilia MTCC 434	MTCC 434	19485	SAMN02952972	PRJNA235261	GCA_000597745.1	◔	4.88156	66.20	-	JALV01	306	4528	4327
Stenotrophomonas maltophilia M30	M30	19485	SAMN02592618	PRJNA235918	GCA_000611735.2	◔	4.90201	66.40	-	JELS02	193	4510	4351
Stenotrophomonas maltophilia SelTE02	SelTE02	35021	SAMEA3138997	PRJEB4721	GCA_000613205.1	◔	4.55711	66.40	-	CBXW01	63	4096	3986
Stenotrophomonas maltophilia	53	19485	SAMN03067892	PRJNA260977	GCA_000758465.1	◔	4.63789	66.30	-	JRJA01	127	4218	4098
Stenotrophomonas maltophilia	B418	19485	SAMN03161950	PRJNA266048	GCA_000788095.1	◔	4.68825	65.50	-	JSXG01	231	4104	3999
Stenotrophomonas maltophilia	ZBG7B	36090	SAMN03280975	PRJNA272355	GCA_000834105.1	◔	4.0654	66.30	-	JXIP01	145	3640	3515
Stenotrophomonas maltophilia	UV74	19485	SAMN03076212	PRJNA261822	GCA_000978875.1	◔	4.88958	66.70	-	LBFT01	179	4528	4359
Stenotrophomonas maltophilia	As1	19485	SAMN03491122	PRJNA272632	GCA_001051925.1	◔	4.39408	66.60	-	LFKU01	33	3858	3720
Stenotrophomonas maltophilia	1030_SMAL	19485	SAMN03196995	PRJNA267549	GCA_001068535.1	◔	4.93283	66.20	-	JWFN01	308	4479	4370
Stenotrophomonas maltophilia	1025_SMAL	19485	SAMN03196989	PRJNA267549	GCA_001069105.1	◔	4.85462	66.30	-	JWFT01	488	4470	4270

Genes	4,576
CDS	4,446
Pseudo Genes	49
rRNAs	3, 1, 1 (5S, 16S, 23S)
complete rRNAs	3, 1, 1 (5S, 16S, 23S)
tRNAs	75
ncRNAs	1

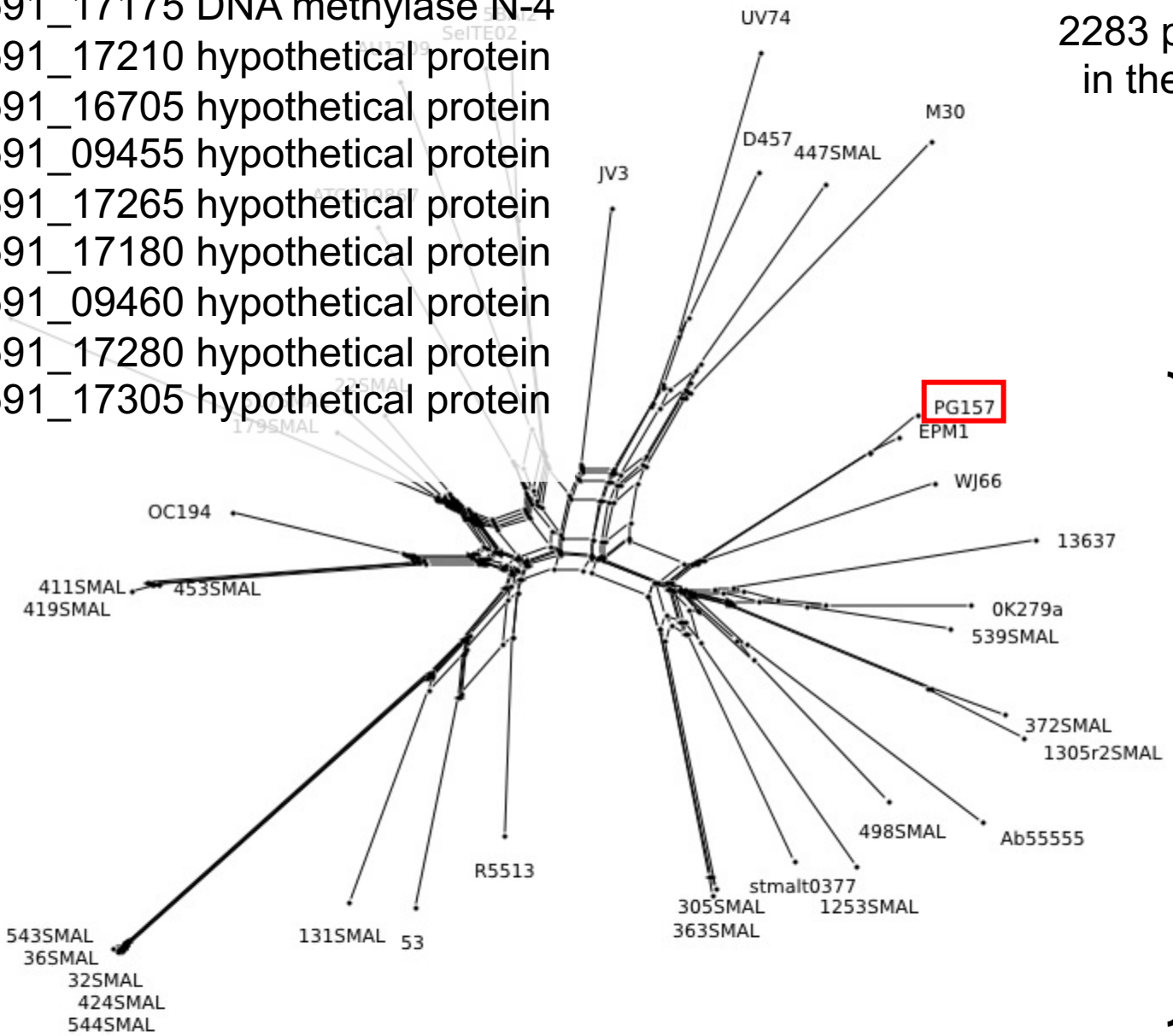
http://www.sanger.ac.uk/science/tools/artemis




5BAI2
SelTE02
AU1209
UV74
M30
D457
447SMAL
JV3
ATCC19867
RR10
22SMAL
517SMAL
179SMAL
PG157
EPM1
WJ66
OC194
13637
411SMAL
453SMAL
419SMAL
0K279a
539SMAL
372SMAL
1305r2SMAL
498SMAL
Ab55555
R5513
stmalt0377
305SMAL
1253SMAL
363SMAL
543SMAL
131SMAL
53
36SMAL
32SMAL
424SMAL
544SMAL

PG 157 Unique proteins (11)
AS591_17215 hypothetical protein
AS591_17330 hypothetical protein
AS591_17175 DNA methylase N-4
AS591_17210 hypothetical protein
AS591_16705 hypothetical protein
AS591_09455 hypothetical protein
AS591_17265 hypothetical protein
AS591_17180 hypothetical protein
AS591_09460 hypothetical protein
AS591_17280 hypothetical protein
AS591_17305 hypothetical protein

2283 proteins (53.39%)
in the core-proteome



This 14 strains share 9 proteins unique of this group